La grande cavità a circa 600 metri di profondità nel sottosuolo della miniera di sale di Morton a est di Cleveland sarà riempita nei prossimi mesi da 10 000 tonnellate d'acqua per preparare una ricerca destinata a rivelare il decadimento di un protone o di un neutrone legato di un nucleo atomico. La cavità ha dimensioni di circa 18×24×21 metri. Il decadimento di ciascuno dei $2{,}5 \times 10^{33}$ protoni e neutroni nella regione centrale dell'acqua darà origine a particelle ad alta velocità. Quando una particella carica attraversa un mezzo trasparente con velocità maggiore di quella della luce in tale mezzo, essa emette un cono di luce azzurra chiamata radiazione Cerenkov. La radiazione è un *bang* ottico analogo al *bang* sonoro generato da un aereo che si muove nell'aria con velocità superiore a quella del suono nello stesso mezzo. La luce Cerenkov verrà rivelata mediante 2400 fotomoltiplicatori, i quali saranno installati lungo le pareti della cavità. L'esperimento viene eseguito sottoterra al fine di ridurre il numero delle particelle di alta energia provenienti dallo spazio che potrebbero essere confuse con prodotti di decadimento di un protone o di un neutrone legato. Il consorzio, al quale saranno affidati la gestione e il funzionamento del rivelatore, comprende ricercatori dell'Università della California a Irvine, dell'Università del Michigan e del Brookhaven National Laboratory.

LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN

numero 156
agosto 1981
anno XIV
volume XXVII

Il decadimento del protone

Il protone ha una vita media pari ad almeno 10^{20} volte l'età attribuita all'universo, ma le teorie indicano che esso non può essere eterno e che tutta la materia ordinaria finirà quindi con il disintegrarsi

di Steven Weinberg

La scoperta della radioattività da parte di Antoine Henri Becquerel nel 1896 smantellò la credenza che tutti gli atomi fossero eterni e immutabili. In seguito si scoprì che le particelle energetiche rivelate da Becquerel erano emesse quando i nuclei degli atomi di una sostanza radioattiva decadono spontaneamente in altri nuclei atomici. Per quanto interessante, tale instabilità nucleare pareva fosse una rarità, una proprietà limitata soltanto a taluni elementi pesanti, quali l'uranio e il radio. I nuclei degli elementi comuni quali l'idrogeno e l'ossigeno, invece, erano ritenuti assolutamente stabili.

Oggi esistono parecchie ragioni teoriche per sospettare che tutti i nuclei atomici finiscano per decadere e che quindi tutta la materia sia in qualche modo radioattiva. In un decadimento di questo genere uno dei due tipi di particelle del nucleo atomico, un protone o un neutrone, verrebbero spontaneamente trasformate in particelle energetiche molto diverse dalle particelle che costituiscono gli atomi ordinari. Persino il nucleo più leggero, quello dell'idrogeno, che è formato da un solo protone, dovrebbe essere soggetto a decadimento.

Un gran numero di prove, a cominciare dalla cospicua età della Terra, indica che la materia non può essere altamente evanescente. Se la materia ordinaria decade, lo fa soltanto molto lentamente, così lentamente che sono necessari esperimenti in scala straordinariamente grande per rivelarne il decadimento. Becquerel scoprì la disintegrazione radioattiva di nuclei di uranio in un cristallo di sali di uranio che pesava forse qualche grammo; per poter osservare la radioattività, molto più debole, associata al decadimento del protone, sarà necessario tenere sotto controllo molte tonnellate di materiale. Ciononostante, sono oggi in corso ricerche sperimentali sul decadimento del protone.

Per vedere che cosa c'è in palio in questi esperimenti, è utile chiederci prima perché nessuna cosa al mondo dovrebbe avere una vita infinita. Per esempio, si pensa ancora che l'elettrone sia assolutamente stabile. Quali principi fisici gli impediscono di decadere in altre particelle? Comprendendo la stabilità di particelle quale l'elettrone si può stabilire se esistono principi fisici che impediscono il decadimento dei nuclei atomici ordinari.

L'esperienza nella fisica delle particelle elementari ci insegna che qualsiasi processo immaginabile di decadimento accadrà spontaneamente a meno che esso non sia vietato da uno dei principi di conservazione della fisica. Un principio di conservazione afferma che il valore totale di una certa grandezza, quale l'energia o la carica elettrica, non può mai cambiare. Anche se un processo di decadimento non viene prodotto direttamente da una delle interazioni fondamentali delle particelle elementari, se esso non è vietato da un principio di conservazione, sarà prodotto da una più o meno complicata successione di emissioni e di assorbimenti di particelle. Quindi nell'esaminare se una particella qualsiasi è stabile ci si deve chiedere se il suo decadimento violerebbe qualche principio di conservazione.

Il principio di conservazione dell'energia è di facile applicazione. Esso richiede semplicemente che la massa della particella che decade (o l'energia equivalente a tale massa) sia maggiore della massa totale dei prodotti di decadimento. (Non basta che le masse siano uguali, perché una certa quantità di massa deve essere trasformata in energia cinetica dei prodotti del decadimento.) Pertanto un buon modo per cominciare l'esame della stabilità di qualsiasi particella è quello di elencare tutte le particelle più leggere in cui essa potrebbe ragionevolmente decadere.

Si consideri l'elettrone. Per quanto ne sappiamo, esistono solo pochi tipi di particelle con massa inferiore a quella dell'elettrone. La più comune di esse è il fotone, il quanto di luce, la cui massa si ritiene sia esattamente nulla. Esistono forti argomentazioni teoriche per credere anche all'esistenza di un quanto di radiazione gravitazionale, il gravitone, ancora con massa nulla. Infine esistono varie specie di particelle chiamate neutrini, simili per certi aspetti all'elettrone; essi vengono emessi nel noto tipo di radioattività chiamato decadimento beta, il tipo scoperto da Becquerel nel 1896. Ai neutrini è stata finora attribuita massa nulla, ma la determinazione della loro massa è oggi l'obiettivo di un intenso sforzo teorico e sperimentale. Ciononostante, non vi sono dubbi che almeno una delle specie di neutrini abbia una massa di circa un decimillesimo della massa dell'elettrone.

Allora, perché l'elettrone non decade, per esempio, in neutrini e fotoni? La ri-

sposta è che anche se tale decadimento soddisferebbe il principio di conservazione dell'energia, esso violerebbe un altro principio di conservazione, quello della carica elettrica. Benjamin Franklin fu il primo a scoprire che la quantità totale di carica elettrica (positiva meno negativa) non aumenta né diminuisce mai, anche se le cariche di segno opposto si possono separare o ricombinare. Gli elettroni trasportano una ben definita carica elettrica negativa, ma tutte le particelle più leggere nelle quali l'elettrone potrebbe decadere (il fotone, il gravitone e i neutrini) trasportano una carica elettrica nulla. Il decadimento di un elettrone implicherebbe la distruzione di una ben definita quantità negativa di carica elettrica e per tale motivo è rigorosamente vietato.

Consideriamo ora come questi principi di conservazione potrebbero essere applicati al decadimento dei due tipi di particelle che costituiscono il nucleo atomico. Per il momento consideriamo soltanto la più leggera delle due particelle, il protone, e ritorneremo più avanti sul neutrone. Il protone trasporta una carica elettrica positiva con lo stesso valore assoluto, ma di segno opposto, a quella dell'elettrone e quindi neppure esso può decadere in neutrini, fotoni o gravitoni. Tuttavia, il protone è circa 1820 volte più pesante dell'elettrone e vi sono numerose particelle di massa inferiore dotate anch'esse di carica positiva. Il protone potrebbe decadere in queste altre particelle senza violare né la conservazione dell'energia né quella della carica elettrica. Per esempio, l'elettrone ha un'antiparticella chiamata positone, con la stessa massa dell'elettrone, ma con una carica elettrica positiva uguale a quella del protone. (Per ogni tipo di particella esiste un'antiparticella con la stessa massa, ma con valori opposti di altre proprietà, quale la carica elettrica. Incidentalmente, il positone è stabile per la stessa ragione per cui lo è l'elettrone.) Non c'è nulla nei principi di conservazione dell'energia o della carica che impedirebbe a un protone di decadere in un positone e in un numero qualsiasi di fotoni e neutrini.

Un altro possibile candidato al ruolo di prodotto di decadimento del protone è l'antimuone. Il muone è una particella simile sotto molti aspetti all'elettrone e con la stessa carica, ma è circa 210 volte più pesante. (Il muone decade in effetti in un elettrone e in neutrini.) L'antimuone ha la stessa carica del protone, ma soltanto circa un nono della sua massa. Un protone potrebbe quindi decadere in un antimuone e in particelle neutre leggere come fotoni e neutrini.

Un ulteriore possibile prodotto del decadimento del protone è un mesone, un membro del gruppo di particelle instabili con massa intermedia tra quella dell'elettrone e quella del protone. I principi di conservazione dell'energia e della carica consentirebbero al protone di decadere, per esempio, in un mesone carico positivamente e in un neutrino o in un mesone neutro e in un positone. Uno qualsiasi di tali processi di decadimento condurrebbe alla distruzione totale dell'atomo di idrogeno. In un elemento più pesante essi modificherebbero la natura chimica dell'elemento e libererebbero energia in quantità molto maggiori di quelle liberate nella radioattività ordinaria.

Per quale motivo non si è mai vista disintegrarsi la materia come risultato di tali processi di decadimento? Pare che questo problema sia stato affrontato per la prima volta da Hermann Weyl nel 1929. Allora positoni, muoni e mesoni erano sconosciuti e quindi gli ipotetici schemi di decadimento del protone sopra indicati non potevano venire immaginati. Ciononostante Weyl aveva dei dubbi sulla stabilità della materia; può darsi che si sia domandato perché i protoni di un atomo non assorbano gli elettroni orbitanti, portando per esempio al decadimento di un atomo di idrogeno in uno sciame di fotoni. Weyl suggerì che la stabilità della materia si potrebbe spiegare se esistessero due tipi di cariche elettriche, una trasportata dall'elettrone e l'altra dal protone. Se ciascuno dei due tipi di carica si conservasse separatamente, la mutua annichilazione di un protone con un elettrone sarebbe vietata. La proposta di Weyl tuttavia non destò molto interesse a quel tempo.

La questione venne ripresa da E.C.G. Stueckelberg nel 1938 e da Eugene P. Wigner (in una nota a pie' di pagina) nel 1949. Essi proposero ciò che doveva diventare il quadro convenzionale, cioè che oltre all'energia e alla carica elettrica c'è un'altra proprietà della materia che si conserva, quella che da allora doveva chiamarsi numero barionico. I barioni (dal greco, *barys*, pesante) sono una famiglia di particelle che comprendono il protone e molte particelle più pesanti del protone, quali il neutrone e particelle altamente instabili chiamate iperoni. A tutti i barioni si assegna un numero barionico +1 mentre tutte le particelle più leggere, tra le quali il fotone, l'elettrone, il positone, il gravitone, il neutrino, il muone e i mesoni hanno un numero barionico nullo. Per un atomo o un altro sistema composto di particelle il numero barionico è la somma dei numeri barionici delle particelle costituenti. Ne consegue che qualunque insieme di particelle più leggere del protone ha numero barionico nullo. Il principio di conservazione del numero barionico è l'affermazione del fatto che il numero barionico totale non può cambiare. Il decadimento di un protone in un insieme di particelle più leggere implicherebbe la trasformazione di uno stato con numero barionico +1 in uno stato con numero barionico nullo, e perciò tale decadimento è vietato.

Un'antiparticella ha un numero barionico opposto a quello della corrispondente particella. L'antiprotone, per esempio, ha numero barionico −1 ed è un antibarione. Un protone e un antiprotone si possono annichilare a vicenda senza violare la conservazione del numero barionico; il protone e l'antiprotone hanno un numero barionico complessivo +1 più −1, cioè zero, e possono pertanto trasformarsi in uno sciame di mesoni o fotoni. Perciò la conservazione del numero barionico non richiede che ogni protone sia eterno, ma piuttosto che i protoni non decadano spontaneamente in materia ordinaria, nella quale non vi sono antiprotoni.

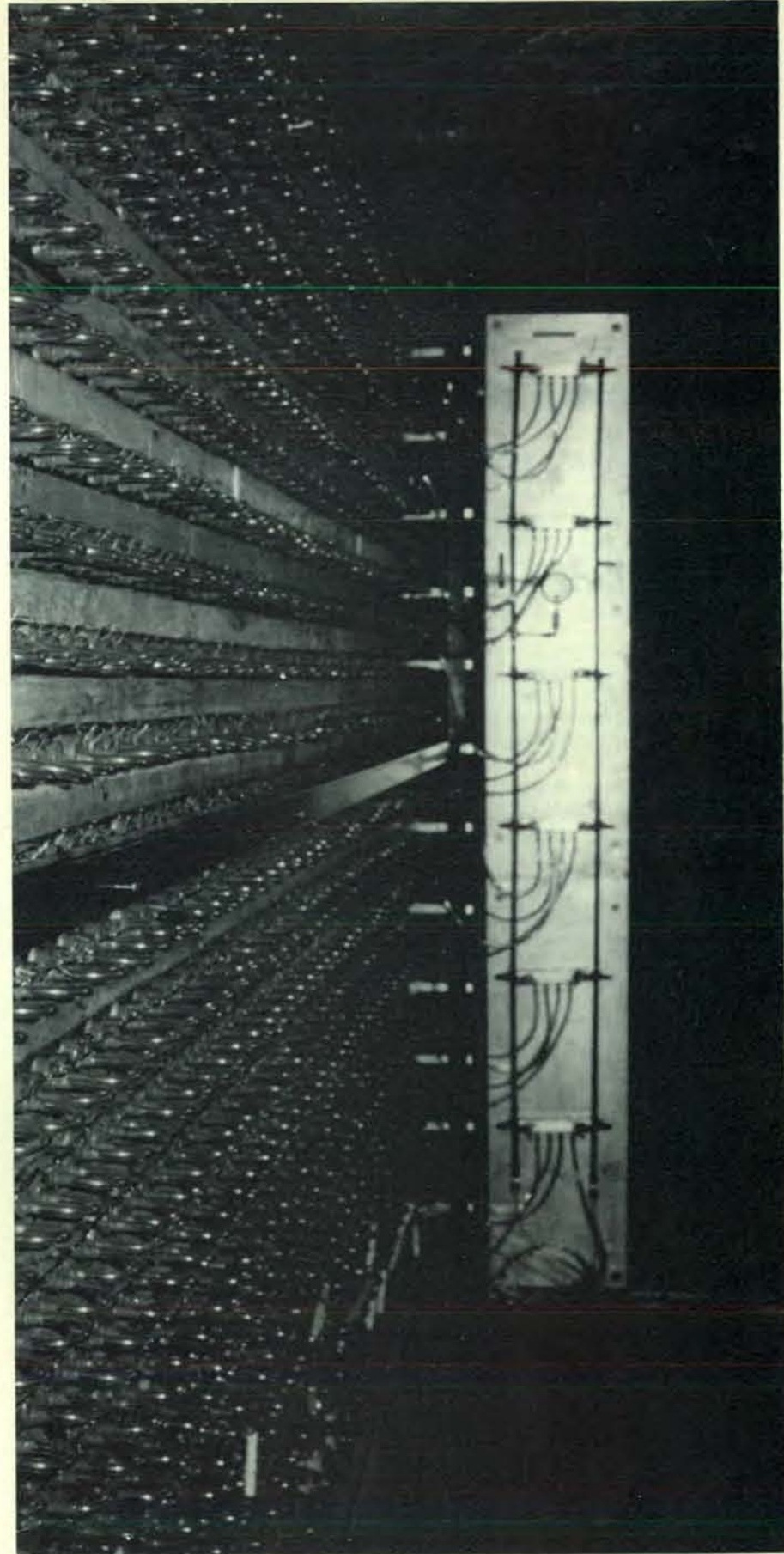
Finora ho esaminato soltanto il decadimento del protone, ma ovviamente il nucleo di molti atomi è costituito non soltanto da protoni, ma anche da neutroni. Quali sono le possibilità di decadimento del neutrone? Il neutrone è un barione con una carica elettrica nulla, ma con una massa leggermente superiore a quella del protone. Per essere più precisi, la massa del neutrone è di poco superiore alla somma della massa del protone e di quella dell'elettrone. Questa relazione suggerisce un possibile modo di decadimento del neutrone: esso potrebbe dare origine a un protone, a un elettrone e a qualche particella neutra priva di massa. Evidentemente in questo processo l'energia si può conservare. Altrettanto accade per la carica elettrica, dal momento che le cariche del protone e dell'elettrone si annullano a vicenda. Anche il numero barionico si conserva, dato che il neutrone e il protone hanno entrambi numero barionico +1 e le altre particelle hanno numero barionico nullo.

Un neutrone libero (cioè non legato in un nucleo atomico) decade proprio in questo modo: esso produce un protone, un elettrone e un antineutrino. La vita media del neutrone libero, cioè il tempo necessario perché la metà dei neutroni in un campione comunque grande decada, è di circa 10 minuti. Anche i neutroni di certi nuclei atomici, quale il nucleo di trizio (l'isotopo pesante dell'idrogeno con un protone e due neutroni), possono decadere in protoni; è il processo del decadimento beta. Tuttavia, nella maggior parte dei nuclei, i neutroni non decadono perché sarebbe necessaria troppa energia per creare un protone tra le forze repulsive elettrostatiche generate dagli altri protoni del nucleo. In tali nuclei i neutroni sono stabili come i protoni.

Resta la possibilità che un neutrone legato in un nucleo possa decadere in qualche altro modo che non conserva il numero barionico. Per esempio, esso potrebbe dare origine a un positone e a un mesone carico negativamente o a un elettrone e a un mesone carico positivamente. La scoperta di un tale decadimento del neutrone sarebbe altrettanto significativa di quella del decadimento di un protone. In realtà gli esperimenti che stanno cercando il decadimento del protone cercano anche il decadimento di neutroni legati. Tuttavia, dal momento che il decadimento dei neutroni liberi è già ben noto, le verifiche sperimentali della conservazione del numero barionico hanno finito con l'essere conosciute come esperimenti sul decadimento del protone.

Negli ultimi anni è diventata opinione diffusa che i barioni e i mesoni siano costituiti dalle particelle più fondamentali chiamate quark. Un barione è formato da tre quark, un antibarione è formato da tre antiquark, mentre un mesone è formato da un quark e da un antiquark. L'elettrone, il muone e i neutrini appartengono alla famiglia di particelle chiamate leptoni, che non sono costituiti da quark e in effetti non presentano tracce di struttura interna. Su questa base il numero barionico di qualsiasi sistema di particelle è esattamente un terzo del numero totale di quark, cioè un terzo della differenza tra il numero di quark e il numero di antiquark. La conservazione del numero complessivo di quark è equivalente alla conservazione del numero barionico.

Il lettore scettico può sentirsi in un certo qual modo insoddisfatto della conservazione del numero barionico come spiegazione della stabilità del protone e del neutrone legato. Dal mio punto di vista la sua sensazione sarebbe giustificata. Il numero barionico è stato inventato come un dispositivo di contabilità, per spiegare la mancata osservazione dei decadimenti del protone e di altri decadimenti correlati; esso non ha altro significato conosciuto. Sotto tale aspetto il numero barionico è molto diverso dalla carica elettrica, che ha un significato dinamico diretto: una carica elettrica crea campi elettrici e magnetici e su essa agiscono a loro volta tali campi, che hanno effetti osservabili sul moto della carica. La teoria dell'elettricità e del magnetismo non avrebbe senso se non si conservasse la carica elettrica, mentre non si conosce alcuno di tali argomenti dinamici per la conservazione del numero barionico.

In realtà, esiste un'evidenza empirica che nega l'esistenza di qualsiasi tipo di campo (chiamiamolo campo baritropico) che possa condurre per il numero barionico alla stessa relazione alla quale conduce il campo elettromagnetico per la carica elettrica. La Terra è formata da circa 4×10^{51} protoni e neutroni e quindi ha un enorme numero barionico. Se la Terra fosse una sorgente di un campo baritropico, ci si dovrebbe aspettare che il campo attragga o respinga i protoni e i neutroni dei corpi ordinari sulla superficie terrestre. Una forza baritropica si potrebbe distinguere dalla forza gravitazionale perché la forza gravitazionale che la Terra esercita su un corpo è proporzionale alla massa del corpo, mentre la forza baritropica dovrebbe essere proporzionale al numero barionico. Corpi di ugual massa composti da elementi differenti possono avere numeri barionici che differiscono di quasi l'1 per cento. Numerosi esperimenti di alta precisione (a partire da quelli di Roland von Eötvös del 1889) mostrano che l'attrazione di corpi da parte della Terra è in realtà rigorosamente proporzionale alla loro massa, non al loro numero barionico. Nel 1955 T. D. Lee della Columbia University e C. N. Yang dell'Institute for Advanced Study di Princeton mostrarono da un'analisi di tali esperimenti che qualsiasi forza baritropica tra due particelle nucleari dovrebbe essere molto più debole della forza gravitazionale, che è a sua volta di cir-

Una lastra di cemento di circa 31 tonnellate nella miniera di Soudan nel Minnesota verrà controllata per eventi che segnalano il decadimento del protone. Nella lastra sono inseriti circa 3450 contatori di particelle, che riveleranno direttamente i prodotti del decadimento. L'esperimento è condotto da ricercatori dell'Università del Minnesota e dell'Argonne National Laboratory.

ca 40 ordini di grandezza più debole della forza elettromagnetica. Non si può escludere in modo assoluto che il numero barionico abbia lo stesso ruolo dinamico della carica elettrica, ma le conclusioni di Lee e di Yang rendono tale ruolo estremamente improbabile.

La conclusione che il numero barionico non ha alcun ruolo dinamico non implica direttamente che il numero barionico non si conservi. In effetti, fin dalla metà degli anni trenta i fisici si sono abituati a un certo numero di altre grandezze che non pare abbiano un significato dinamico come quello della carica elettrica e pur tuttavia si conservano, almeno in certi contesti. Tra queste grandezze vi sono quelle chiamate stranezza, isospin e coniugazione di carica. Per fare un esempio, ai protoni e ai neutroni si assegna stranezza nulla, ad alcuni iperoni una stranezza −1 e ad alcuni mesoni chiamati mesoni *K* una stranezza +1. La conservazione della stranezza è stata introdotta come una regola di contabilità per spiegare l'osservazione che un mesone *K* o un iperone non si possono produrre singolarmente in collisioni di nuclei atomici ordinari, mentre si possono produrre in coppia, perché un mesone *K* e un iperone hanno una stranezza totale nulla. Per anni, dopo l'introduzione dell'idea del numero barionico, apparve plausibile che la conservazione del numero barionico fosse un'altra di tali regole non dinamiche di contabilità, che risulta essere universalmente rispettata.

Questo modo di considerare i principi di conservazione è stato radicalmente mutato dallo sviluppo delle moderne teorie delle interazioni tra particelle elementari. Le teorie descrivono tutte le forze conosciute tra le particelle elementari (esclusa la gravitazione) in modo molto simile a quello in cui l'elettromagnetismo fu descritto nella vecchia teoria delle interazioni puramente elettromagnetiche; l'ultima teoria, chiamata elettrodinamica quantistica, fu sviluppata tra gli anni trenta e gli anni quaranta. Oggi si ritiene che esistano 12 campi simili al campo elettromagnetico dell'elettrodinamica quantistica. Si tratta degli otto campi dei gluoni che forniscono le forze nucleari forti che tengono insieme i quark all'interno dei barioni e dei mesoni e dei quattro campi elettrodeboli che forniscono in modo unificato sia le forze nucleari deboli responsabili del decadimento beta sia lo stesso elettromagnetismo. Esistono in corrispondenza 12 principi di conservazione, simili alla conservazione della carica elettrica, per grandezze chiamate colore, isospin elettrodebole e ipercarica elettrodebole. (Il colore è una proprietà dei quark che nulla ha a che vedere con i colori visibili; la carica elettrica è una particolare combinazione pesata dell'ipercarica elettrodebole e dell'isospin elettrodebole.) Diversamente dal numero barionico, queste grandezze che si conservano hanno un diretto significato fisico: sono le particelle che trasportano tali grandezze a dare origine ai campi gluonici e ai campi elettrodeboli, e tali campi esercitano a loro volta una forza su ciascuna di tali particelle. La forza dipende dai valori delle 12 grandezze che sono trasportate dalla particella.

Mentre apparivano i nuovi principi di conservazione, i vecchi principi di conservazione non dinamici sono stati in un certo senso demoliti. Per esempio, la moderna teoria delle interazioni nucleari forti è così strettamente vincolata dalla conservazione del colore (e da altri principi) da non poter in alcun modo comprendere i tipi di complicazioni che sarebbero necessari per violare la conservazione della stranezza. Si potrebbe cercare di introdurre interazioni fondamentali che non conservano la stranezza, ma risulta sempre che si può ridefinire che cosa si intende per stranezza in modo tale che essa si conservi ancora. La conservazione della stranezza è perciò interpretata oggi non come un principio fondamentale quale la conservazione dell'energia o la conservazione della carica, ma come una conseguenza della teoria dettagliata delle interazioni forti e in particolare del principio genuinamente fondamentale della conservazione del colore. Dal momento che la conservazione della stranezza non è un principio fondamentale della fisica, non c'è alcun motivo generale perché essa venga rispettata al di fuori del regno delle forze forti. In realtà, sin dalla scoperta della stranezza, è risaputo che essa non si conserva per le forze nucleari deboli.

Anche gli altri principi di conservazione non dinamici hanno subito un'analoga demolizione di fatto; essi non vengono più visti come principi di conservazione fondamentali, allo stesso livello di quelli di conservazione dell'energia o della carica, ma piuttosto come semplici conseguenze matematiche della struttura delle attuali teorie delle interazioni tra le particelle elementari. Un elenco dei principi di conservazione che oggi paiono fondamentali comprenderebbe le 12 grandezze associate alle forze forte ed elettrodebole, la conservazione di grandezze quali l'energia e la quantità di moto che sono associate allo stesso modo alle forze gravitazionali e la conservazione del numero barionico, che non si sa se è associata a una forza qualsiasi.

Questo fatto da solo dovrebbe farci dubitare della conservazione del numero barionico: il numero barionico non ha alcuna necessità di conservarsi allo stesso modo in cui si devono conservare altre grandezze quali l'energia, la carica, il colore e simili per poter disporre di teorie accettabili delle interazioni tra particelle elementari. Inoltre, esistono indicazioni chiare del fatto che la conservazione del numero barionico non è esatta. Una delle indicazioni viene data dalla moderna teoria delle interazioni elettrodeboli. Gerard't Hooft dell'Università di Utrecht ha dimostrato che in questa teoria certi piccoli effetti che non si possono rappresentare con alcun numero finito di emissioni e di assorbimenti di particelle elementari portano a processi che non conservano i barioni, ma tali processi sono straordinariamente poco frequenti. Tali processi sono troppo lenti per essere rivelati, ma è interessante che essi nascano proprio perché la conservazione del numero barionico non è correlata ad alcun tipo di campo baritropico; nessuno di tali effetti potrebbe avere come conseguenza la mancata conservazione di grandezze, quali la carica elettrica, correlate a campi di forza.

Un'altra indicazione della possibilità che il numero barionico non si conservi ci viene dalla cosmologia. Sulla base di considerazioni anche solo estetiche, si sarebbe potuto supporre che l'universo abbia avuto inizio con uguali quantità di materia e di antimateria e quindi con ugual numero di barioni e di antibarioni. Secondo questa ipotesi l'universo potrebbe aver avuto inizio con un numero barionico totale uguale a zero. Se il numero barionico dovesse conservarsi, il numero barionico totale dovrebbe rimanere nullo. Quasi tutti i protoni e i neutroni avrebbero dovuto annichilarsi con collisioni con antiprotoni e antineutroni e oggi l'universo dovrebbe contenere soltanto un sottile impasto di fotoni e neutrini, senza stelle né pianeti né scienziati.

È possibile che l'universo abbia avuto inizio con un eccesso di materia sull'antimateria, in modo da lasciare indietro qualcosa dopo l'annichilazione di particelle e antiparticelle. È anche possibile (anche se solitamente considerato improbabile) che la materia e l'antimateria si siano in qualche modo separate e che noi viviamo in una regione con numero barionico positivo in un universo il cui numero barionico totale sia nullo. Però, se il numero barionico non si conserva, esiste una possibilità più interessante, e cioè che l'universo abbia avuto inizio con uguali quantità di materia e di antimateria e che l'attuale eccesso di particelle con numero barionico positivo sia dovuto a processi fisici che hanno violato la conservazione del numero barionico. (È risaputo fin dall'esperimento condotto nel 1964 da James H. Christenson, James W. Cronin, Val L. Fitch e René Turlay della Princeton University che non c'è alcuna esatta simmetria materia-antimateria che richieda processi con creazione di antibarioni con la stessa frequenza di quelli con creazione di barioni.) Tali considerazioni, insieme all'assenza di forze baritropiche evidenziata da Lee e Yang, hanno indotto alcuni teorici (tra i quali il fisico sovietico Andrei D. Sakharov e me) a suggerire negli anni sessanta che il numero barionico possa non conservarsi esattamente. Anche considerazioni cosmologiche hanno ispirato almeno uno degli esperimenti sul decadimento del protone eseguiti in questo periodo, quello di T. Alväger, I. Martinson e H. Ryde dell'Università di Stoccolma e dell'Istituto Nobel. Negli ultimi anni molti teorici hanno elaborato schemi per la produzione di barioni nell'universo primordiale.

Qualunque ipotesi sulla possibile non conservazione del numero barionico deve essere immediatamente messa a confronto col fatto che la materia ordinaria è stabile. Maurice Goldhaber del Brookhaven National Laboratory ha sottolineato che «sappiamo per esperienza» che la vita media del protone è maggiore di 10^{16} anni circa. Se la vita media fosse appena più breve, i circa 10^{28} protoni del corpo umano decaderebbero con una frequenza media di più di 10^{12} protoni all'anno, pari a 30 000 decadimenti al secondo e saremmo tutti in pericolo di vita.

Naturalmente si può porre un limite più rigoroso alla vita media del protone ricercando attivamente il decadimento del protone. Il primo esperimento di questo tipo è stato eseguito nel 1954 da Frederick Reines e Clyde L. Cowan, Jr., allora presso il Los Alamos Scientific Laboratory, e da Goldhaber. Essi impiegarono circa 300 litri di uno scintillatore a idrocarburi, un materiale in cui le particelle cariche energetiche prodotte da un decadimento di un protone generano un lampo di luce rivelabile. Come in tutti i successivi esperimenti sul decadimento del protone, l'apparecchiatura venne posta sottoterra per schermarla dai raggi cosmici. (Le particelle energetiche di tali raggi possono provocare eventi che potrebbero essere interpretati come decadimenti di un protone.) Con tali precauzioni essi osservarono soltanto alcune scintillazioni al secondo, quasi tutte attribuibili a raggi cosmici che penetravano in profondità nel sottosuolo. Reines, Cowan e Goldhaber conclusero che la vita media di un protone o di un neutrone legato deve essere maggiore di circa 10^{22} anni.

Successivi esperimenti eseguiti da numerosi fisici hanno gradualmente aumentato il limite inferiore empirico della vita media del protone. La più elaborata ricerca fin qui condotta, i cui risultati

PARTICELLA	MASSA (MeV)	CARICA ELETTRICA	NUMERO BARIONICO	PRINCIPALE MODO DI DECADIMENTO
FOTONE (γ)	0	0	0	NESSUN MODO CONOSCIUTO
NEUTRINO (ν)	0?	0	0	NESSUN MODO CONOSCIUTO
ANTINEUTRINO ($\bar{\nu}$)	0?	0	0	NESSUN MODO CONOSCIUTO
ELETTRONE (e^-)	0,511	−1	0	NESSUN MODO CONOSCIUTO
POSITONE (e^+)	0,511	+1	0	NESSUN MODO CONOSCIUTO
MUONE (μ^-)	105,7	−1	0	$\mu^- \to e^- + \nu + \bar{\nu}$ MASSA: 105,7 → 0,511 + 0 + 0 CARICA: −1 → −1 + 0 + 0 NUMERO BARIONICO: 0 → 0 + 0 + 0
ANTIMUONE (μ^+)	105,7	+1	0	$\mu^+ \to e^+ + \nu + \bar{\nu}$ MASSA: 105,7 → 0,511 + 0 + 0 CARICA: +1 → +1 + 0 + 0 NUMERO BARIONICO: 0 → 0 + 0 + 0
PIONI O MESONI PI (π^+)	139,6	+1	0	$\pi^+ \to \mu^+ + \nu$ MASSA: 139,6 → 105,7 + 0 CARICA: +1 → +1 + 0 NUMERO BARIONICO: 0 → 0 + 0
(π^0)	135	0	0	$\pi^0 \to \gamma + \gamma$ MASSA: 135 → 0 + 0 CARICA: 0 → 0 + 0 NUMERO BARIONICO: 0 → 0 + 0
(π^-)	139,6	−1	0	$\pi^- \to \mu^- + \bar{\nu}$ MASSA: 139,6 → 105,7 + 0 CARICA: −1 → −1 + 0 NUMERO BARIONICO: 0 → 0 + 0
MESONI *K* (K^+)	493,7	+1	0	$K^+ \to \mu^+ + \nu$ MASSA: 493,7 → 105,7 + 0 CARICA: +1 → + 1+ 0 NUMERO BARIONICO: 0 → 0 + 0
(K^0_S)	497,7	0	0	$K^0_S \to \pi^+ + \pi^-$ MASSA: 497,7 → 139,6 + 139,6 CARICA: 0 → +1 + −1 NUMERO BARIONICO: 0 → 0 + 0
(K^0_L)	497,7	0	0	$K^0_L \to \pi^0 + \pi^0 + \pi^0$ MASSA: 497,7 → 135 + 135 + 135 CARICA: 0 → 0 + 0+ 0 NUMERO BARIONICO: 0 → 0 + 0 + 0
(K^-)	493,7	−1	0	$K^- \to \mu^- + \bar{\nu}$ MASSA: 493,7 → 105,7+ 0 CARICA: −1 → −1 + 0 NUMERO BARIONICO: 0 → 0 + 0
PROTONE (p)	938,3	+1	+1	NESSUN MODO CONOSCIUTO
ANTIPROTONE ($\bar{p}$)	938,3	−1	−1	NESSUN MODO CONOSCIUTO
NEUTRONE (n)	939,6	0	+1	$n \to p + e^- + \bar{\nu}$ MASSA: 939,6 → 938,3 +0,511 + 0 CARICA: 0 →+1 + −1 + 0 NUMERO BARIONICO: +1 → +1 + 0 + 0
ANTINEUTRONE ($\bar{n}$)	939,6	0	−1	$\bar{n} \to \bar{p} + e^+ + \nu$ MASSA: 939,6 → 938,3 +0,511 + 0 CARICA: 0 → −1 + +1 + 0 NUMERO BARIONICO: −1 → −1 + 0+ 0
IPERONE Λ (Λ^0)	1115,6	0	+1	$\Lambda^0 \to p + \pi^-$ MASSA: 1115,6 → 938,3 + 139,6 CARICA: 0 → +1 + −1 NUMERO BARIONICO: +1 → +1 + 0
ANTIIPERONE Λ ($\bar{\Lambda}^0$)	1115,6	0	−1	$\bar{\Lambda}^0 \to \bar{p} + \pi^+$ MASSA: 1115,6 → 938,3 + 139,6 CARICA: 0 → −1 + +1 NUMERO BARIONICO: −1 → −1 + 0

I principali modi di decadimento delle particelle sono retti da principi di conservazione fondamentali. Il principio di conservazione dell'energia richiede che la massa della particella che decade sia almeno pari alla massa totale dei prodotti di decadimento. Il principio di conservazione della carica elettrica richiede che la carica della particella che decade sia uguale alla carica totale dei prodotti di decadimento. Il principio di conservazione del numero barionico afferma che il numero barionico della particella che decade deve uguagliare la somma dei numeri barionici dei prodotti di decadimento. Il protone e molte particelle più pesanti del protone hanno numero barionico +1. Tutte le particelle più leggere del protone hanno un numero barionico nullo. Le particelle sono elencate in ordine crescente di massa. La massa è espressa in termini dell'energia equivalente, in unità di un milione di elettronvolt (MeV). La carica è espressa in unità di carica del protone. Sotto ciascun modo di decadimento è riportato un rendiconto delle tre grandezze che si conservano. Il decadimento di un protone in un insieme di particelle più leggere implicherebbe una trasformazione di uno stato con numero barionico +1 in uno con numero barionico nullo, e pertanto il decadimento è vietato dalla conservazione del numero barionico. Se si scoprirà che il protone decade, ciò significa che il principio di conservazione del numero barionico non è universalmente rispettato.

sono stati pubblicati, è stata eseguita da un consorzio di ricercatori della Case Western Reserve University, dell'Università di Witwatersrand e dell'Università della California a Irvine. Essi hanno posto sotto controllo 20 tonnellate di uno scintillatore a idrocarburi a una profondità di 3200 metri in una miniera d'oro del Sud Africa dal 1964 al 1971. Una recente analisi dei loro dati ha portato al risultato che la vita media di un protone o di un neutrone legato è superiore a circa 10^{30} anni.

Si tratta veramente di una lunga vita media. A scopo di confronto l'attuale età dell'universo è valutata in soli 10^{10} anni. Si può sperare di osservare il decadimento di particelle con vite medie così lunghe solo perché i processi di decadimento radioattivo si manifestano statisticamente: un campione di particelle con una vita media di t anni non sopravvive per t anni e non tutte le particelle decadono all'unisono: una frazione $1/t$ del numero totale di particelle decadrà nel primo anno, $1/t$ delle particelle restanti decadranno nell'anno successivo, e così via. L'estremo inferiore della vita media del protone viene imposto non osservando un solo protone per lungo tempo e aspettando che esso decada, ma osservando i 10^{31} protoni e neutroni delle 20 tonnellate di scintillatore per diversi anni e aspettando che qualche dozzina di essi decada.

È la lunga vita media del protone che ha portato all'idea della conservazione del numero barionico. Come potrebbe il protone vivere così a lungo se non ci fosse qualche principio di conservazione che lo fa vivere in eterno? Negli ultimi anni si è avuta una risposta.

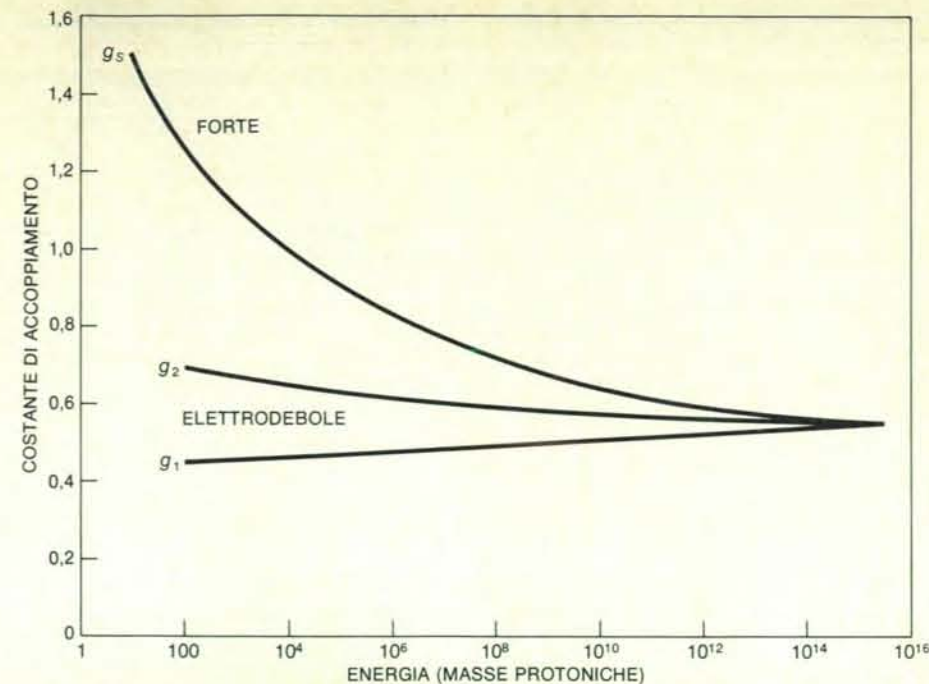
Ricordiamo che la moderna teoria delle interazioni deboli, elettromagnetica e forte è altamente vincolata, al punto tale che, per esempio, è impossibile che le interazioni forti violino il principio di conservazione della grandezza chiamata stranezza. La teoria è così vincolata da non poter essere abbastanza complicata (a parte il piccolissimo effetto 't Hooft) da permettere qualsiasi violazione della conservazione del numero barionico, a meno che non si introducano nuovi tipi di particelle con valori esotici della carica, del colore e così via. Tali particelle dovrebbero essere qualitativamente differenti da qualsiasi particella oggi conosciuta.

Il decadimento del protone con l'emissione e l'assorbimento di una particella esotica pesante condurrebbe con molta probabilità a un positone (e^+) e a un mesone neutro (π^0) oppure a un antineutrino ($\bar{\nu}$) e a un mesone positivo (π^+). Si pensa che il protone sia una particella composta, e che nella sua composizione entrino tre costituenti chiamati quark; un mesone è invece formato da un quark e da un antiquark. L'emissione e l'assorbimento di una particella esotica hanno l'effetto di trasformare due quark in un antiquark e in un positone *(le quattro reazioni in alto)* oppure in un antiquark e in un antineutrino *(le tre reazioni in basso)*. I quark, i positoni, gli antineutrini e le particelle esotiche sono indicati dalla loro carica elettrica. La carica si conserva in sette reazioni, che rappresentano in pratica tutti i modi in cui l'emissione e l'assorbimento di una sola particella esotica potrebbero portare al decadimento di un protone in un mesone e in un positone o in un mesone e in un antineutrino. Molti tipi di particelle esotiche potrebbero portare al decadimento del protone. Esse sono distinte dalla loro carica elettrica e dal loro momento angolare intrinseco di spin. Lo spin può essere uguale a zero (come per il mesone pi) oppure a 1 (come per il fotone).

Se si ammette l'esistenza di particelle esotiche del tipo giusto, il decadimento del protone diventa possibile. I principi familiari di conservazione della carica, del colore e così via indicano che è necessaria una particella con una carica elettrica di +4/3, +1/3 o −2/3 per la carica del protone; la particella deve anche avere un momento angolare intrinseco di spin uguale a 0 o a 1 e un colore identico a quello dell'antiquark. Per esempio, una siffatta particella esotica potrebbe venire prodotta quando un quark si trasforma in un antiquark e potrebbe venire distrutta quando un altro quark si trasforma in un antileptone (un positone, un antimuone o un antineutrino); in tal modo i tre quark che formano il protone potrebbero decadere in un antileptone e in un mesone formato dal quark residuo e dall'antiquark.

Tutte queste particelle esotiche dovrebbero essere molto pesanti, altrimenti sarebbero già state rivelate. Se fossero sufficientemente pesanti, esse verrebbero emesse e riassorbite solo con difficoltà e quindi indurrebbero soltanto una piccolissima frequenza di decadimenti del protone. In questo modo è oggi possibile spiegare la lunga vita media del protone senza ipotizzare alcun fondamentale principio di conservazione che assicuri la sua vita eterna, e ciò lascia aperta la possibilità che esso non viva in eterno.

Quanto dovrebbero essere pesanti le particelle esotiche per poter spiegare la lunga vita media del protone? Supponendo che le particelle esotiche interagiscano più o meno come i fotoni, si può calcolare approssimativamente che una vita media del protone superiore a 10^{30} anni richiede una particella esotica la cui massa sia superiore a circa 10^{14} masse protoniche. Si tratta di una massa spaventosamente grande, superiore a quella che si possa mai sperare di produrre con qualsiasi acceleratore che si possa oggi progettare. Tuttavia ci sono almeno due motivi per sospettare che tali particelle enormemente pesanti possano davvero esistere.

Il primo motivo ha a che vedere con il fenomeno della gravitazione, che è stato finora lasciato fuori da queste considerazioni. La teoria generale della relatività di Einstein fornisce una soddisfacente teoria delle interazioni gravitazionali delle particelle a tutte le energie sperimentalmente accessibili. Tuttavia, a causa delle fluttuazioni quantistiche, la teoria non è più valida ad altissime energie, dell'ordine di 10^{19} masse protoniche. Tale valore è noto come massa di Planck, da Max Planck che nel 1900 notò che una tale massa dovrebbe comparire naturalmente in qualsiasi tentativo di combinare la sua teoria dei quanti con la teoria della gravitazione. La massa di Planck è all'incirca l'energia alla quale la forza gravitazionale tra particelle diventa più intensa di quella elettrodebole o della forza forte. Per poter evitare un'incoerenza tra la meccanica quantistica e la relatività generale, devono entrare nella fisica alcuni nuovi aspetti, a una certa energia inferiore o uguale a 10^{19} masse protoniche.

L'intensità delle tre forze fondamentali (la forza forte, la forza elettromagnetica e la forza debole) che regolano le interazioni delle particelle elementari varia con l'energia dei processi nei quali vengono misurate le forze. La forza elettromagnetica agisce tra particelle dotate di carica elettrica, la forza forte tiene legati insieme i quark per formare i protoni e certe altre particelle, mentre la forza debole è responsabile di certi decadimenti radioattivi. La forza elettromagnetica e la forza debole sono state associate in un'unica teoria chiamata teoria elettrodebole, che ipotizza due tipi di forza elettrodebole. L'intensità intrinseca della forza forte e delle forze elettrodeboli è data da tre costanti di accoppiamento adimensionali (indicate g_s per la forza forte e g_1 e g_2 per le forze elettrodeboli). Anche se tali quantità sono chiamate costanti, esse variano lentamente con l'energia. Se le forze hanno un'origine comune, si prevede che per un certo valore dell'energia le costanti di accoppiamento abbiano lo stesso valore; in altre parole, ci si aspetta che le tre forze abbiano la stessa intensità. Nell'ambito di assunzioni piuttosto generali, si è calcolata un'energia di unificazione delle interazioni forti e di quelle elettrodeboli pari a circa 10^{15} masse protoniche.

L'altro motivo per aspettarci che compaiano nuovi gradi di libertà a energie grandissime ha a che fare con le interazioni elettrodebole e forte. La moderna teoria di tali interazioni coinvolge tre parametri noti come costanti di accoppiamento. Una delle costanti di accoppiamento, chiamata g_s, descrive l'intensità con cui i campi gluonici della forza forte interagiscono con particelle che trasportano le grandezze conservate chiamate colore; le altre due costanti di accoppiamento, chiamate g_1 e g_2, descrivono l'intensità con la quale i campi elettrodeboli interagiscono con particelle che trasportano le corrispondenti grandezze, ipercarica elettrodebole e isospin elettrodebole. Sarebbe bello pensare che tutte queste interazioni abbiano qualche origine comune, nel qual caso le costanti di accoppiamento dovrebbero avere tutte lo stesso ordine di grandezza. Ma ciò è in apparente contrasto con il fatto evidente che le interazioni forti sono forti; le misurazioni forniscono un valore di g_s molto maggiore dei valori di g_1 e g_2.

Una soluzione del problema fu proposta alla Harvard University nel 1974 da Howard Georgi, Helen R. Quinn e da me. Era risaputo fin dal 1954 con la ricerca di Murray Gell-Mann del California Institute of Technology e di Francis E. Low del Massachusetts Institute of Technology che le costanti di accoppiamento dipendono sensibilmente dall'energia dei processi fisici nei quali vengono misurate. Nel 1973 calcoli indipendenti di H. David Politzer di Harvard e David Gross e Frank Wilczek di Princeton mostrarono che la costante di accoppiamento forte, g_s, diminuisce lentamente all'aumentare dell'energia. Anche la maggiore delle due costanti di accoppiamento elettrodeboli, g_2, diminuisce, ma più lentamente, mentre la costante di accoppiamento elettrodebole più piccola, g_1, aumenta con l'aumentare dell'energia. Ciò che Georgi, Quinn e io abbiamo proposto era che la scala di energie alla quale le interazioni forti si unificano con le interazioni elettrodeboli è enormemente alta, tanto alta che la lentissima diminuzione con l'aumentare dell'energia della costante di accoppiamento forte e persino la ancor più lenta variazione con l'energia delle due costanti di accoppiamento elettrodeboli le porta tutte praticamente allo stesso valore a questa elevatissima scala dell'energia. Per essere precisi, sulla base di ipotesi piuttosto generali (come, in pratica, il fatto che le interazioni forte ed elettrodebole vengano unificate da qualche gruppo di simmetria conosciuto matema-

ticamente come gruppo «semplice», che non esistano stadi di unificazione intermedi e che le particelle elementari con spin semiintero formino schemi più o meno simili ai familiari schemi dei leptoni e dei quark) abbiamo scoperto che la scala energetica alla quale le interazioni forte ed elettrodebole vengono unificate è dell'ordine di valori compresi tra 10^{15} e 10^{16} masse protoniche.

Qualsiasi teoria che unifica le interazioni forte ed elettrodebole e pone sullo stesso piano leptoni e quark non dovrebbe coinvolgere nuovi tipi di particelle per completare questo quadro e, come ho sottolineato prima, non esiste alcuna buona ragione per credere che le interazioni di tali nuove particelle conservino il numero barionico. (La scala energetica da 10^{15} a 10^{16} masse protoniche calcolata da Georgi, Quinn e da me è sufficientemente alta, tanto che le interazioni che non conservano il numero barionico e che potrebbero essere prodotte da particelle esotiche così pesanti non condurrebbero a una vita media del protone in contraddizione con l'attuale limite sperimentale inferiore di 10^{30} anni. Abbiamo calcolato una vita media che è, molto approssimativamente, dell'ordine di 10^{32} anni.)

A partire dal 1973 molti teorici si sono dedicati allo sviluppo di tali teorie, e tra essi Jogesh C. Pati dell'Università del Maryland e Abdus Salam del Centro internazionale di fisica teorica di Trieste, Georgi e Sheldon Lee Glashow di Harvard, Harold Fritsch e Peter Minkowski del Cal Tech e Feza Gürsey, Pierre Ramond e Pierre Sikivie della Yale University. I modelli sono generalmente noti col nome del gruppo matematico di simmetrie che correla le varie forze, quali $SU(4)^4$, SU(5), SO (10), E_6, E_7 e SU(7). Tutti questi modelli comprendono particelle esotiche che quando vengono emesse o assorbite trasformano un quark in un antiquark, un leptone in un antileptone; pertanto, come già sottolineato da Pati e Salam nella prima comunicazione sull'unificazione delle interazioni forte ed elettrodebole, esse possono violare la conservazione del numero barionico. Inoltre, tutti questi modelli, in qualche versione almeno, soddisfano alle ipotesi generali avanzate da Georgi, da Quinn e da me, in modo tale che la scala di masse delle particelle esotiche è prevista dell'ordine di 10^{15}-10^{16} masse protoniche e la vita media del protone dell'ordine di 10^{32} anni, secondo la nostra stima approssimata.

Calcoli perfezionati più recenti sono stati eseguiti da molti teorici, tra i quali Andrzej Buras, John Ellis, Mary K. Gaillard e Demetres V. Nanopoulos dell'Organizzazione europea per le ricerche nucleari (CERN), Terrence J. Goldman e Douglas A. Ross del Cal Tech, William J. Marciano e Alberto Sirlin della Rockefeller University e della New York University, Cecilia Jarlskog e Francesco Yndurain del CERN e Lawrence Hall di Harvard. I calcoli più recenti forniscono un valore più attendibile di circa 10^{15} masse protoniche per la scala di masse superpesanti e una vita media del protone di circa 10^{31} anni. Sfortunatamente il calcolo della frequenza di decadimento del protone è complicato dalla presenza delle forze nucleari forti agenti tra i quark e gli antiquark del protone e dei prodotti di decadimento e, quindi, anche se si conoscessero esattamente le proprietà delle particelle esotiche superpesanti, sarebbe probabilmente impossibile prevedere la vita media del protone con precisione superiore a circa un ordine di grandezza.

Un rivelatore da 1000 tonnellate è in costruzione nella miniera di Silver King presso Park City, nello Utah. Il rivelatore comprenderà 800 fotomoltiplicatori distribuiti in un volume d'acqua. I fotomoltiplicatori sono sostenuti da cappi di filo e la camera è circondata da altri contatori di particelle e da uno spesso schermo in cemento. Nel volume d'acqua controllato saranno contenuti circa 6×10^{32} tra protoni e neutroni. Il rivelatore è in fase di costruzione ad opera di ricercatori della Harvard University, della Purdue University e dell'Università del Wisconsin.

Gli studi sperimentali compiuti sulle interazioni deboli hanno già dato una limitata verifica dell'analisi generale eseguita da Georgi, Quinn e da me. Non dovrebbe destare sorpresa il fatto di trovare che i grafici di due qualsiasi delle tre costanti di accoppiamento si intersecano in qualche punto, ma perché le tre curve delle costanti di accoppiamento in funzione dell'energia possano incrociarsi nello stesso punto è necessario imporre una condizione sui loro punti di partenza, cioè sui valori delle costanti di accoppiamento a basse energie. Abbiamo impiegato questa condizione per calcolare che, nelle nostre ipotesi generali, un certo parametro (correlato al rapporto tra g_1 e g_2), che descrive l'unificazione delle interazioni debole ed elettromagnetica, ha un valore prossimo a 0,2. Esperimenti di fisica delle interazioni tra elettroni e neutrini indicano correntemente un valore di tale parametro compreso tra 0,2 e 0,23. I valori teorico e sperimentale sono abbastanza vicini da incoraggiarci a prendere sul serio questa analisi, anche se essa implica un'estrapolazione mai in precedenza verificatasi: 13 ordini di grandezza in energia.

Una massa dell'ordine di 10^{15} masse protoniche è talmente grande che l'emissione e l'assorbimento di particelle così pesanti è quasi impossibile alle energie accessibili sperimentalmente e può pertanto produrre solo effetti piccolissimi in qualsiasi esperimento realizzabile. La sola speranza di poter rivelare tali piccolissimi effetti sta nella possibilità che essi possano violare principi di conservazione altrimenti esatti e rendere in tal modo possibili processi che sarebbero altrimenti rigorosamente vietati. Uno di tali principi di conservazione è la conservazione del numero barionico, che viene verificata cercando il decadimento del protone. L'unico altro principio di conservazione non richiesto per la coerenza delle teorie delle interazioni di particelle e che potrebbe perciò essere violato da effetti a energia altissima è la conservazione del numero leptonico, la conservazione del numero totale di leptoni (neutrini, elettroni, muoni e così via) meno il numero di antileptoni. La mancata conservazione del numero leptonico potrebbe manifestarsi in processi quali il doppio decadimento beta senza neutrini: il decadimento di due protoni di un nucleo in due neutroni più due positoni. In questa reazione il numero barionico rimarrebbe costante, ma il numero leptonico diminuirebbe di 2. (La reazione non viola la conservazione dell'energia perché i neutroni nello stato finale non sono liberi, ma hanno un'energia di legame negativa nel nucleo.) Un altro segno della violazione della conservazione del numero leptonico sarebbe una massa neutrinica non nulla.

Sono oggi in corso numerosi nuovi tentativi per determinare la vita media del protone. Esperimenti sono già in corso nella miniera di Soudan nel Minnesota, nel campo aurifero di Kolar nell'India meridionale, nel tunnel del Monte Bianco tra la Francia e l'Italia e nella valle di Baksan nella regione del Caucaso in Unione Sovietica. Entro la fine dell'anno altri tre esperimenti dovrebbero raccogliere dati; essi sono in corso nella miniera di sale di Morton nell'Ohio, nella miniera Silver King nell'Utah e in un altro punto del tunnel del Monte Bianco. Altri esperimenti sono allo studio.

La radiazione Cerenkov è un segnale tipico non soltanto del decadimento del protone, ma anche di interazioni di particelle ad alta energia provenienti dallo spazio. In effetti, la maggior parte dei lampi di luce rivelati dai fotomoltiplicatori non saranno originati dai prodotti di decadimento del protone. Il numero dei coni di radiazione Cerenkov e la loro orientazione renderanno possibile distinguere il decadimento del protone da altri eventi, quale la collisione di un neutrino creato nell'atmosfera terrestre dai raggi cosmici con un protone o un neutrone del rivelatore. Il decadimento di un protone in un positone e in un pione (o mesone pi) neutro, per esempio, potrebbe dare origine a tre coni di luce Cerenkov a certi angoli caratteristici. Il positone genererebbe un cono, mentre il decadimento del pione neutro in due fotoni, ciascuno dei quali potrebbe dare origine a uno sciame di particelle elettricamente cariche, dovrebbe essere responsabile degli altri due coni di luce. L'interazione del neutrino produce soltanto una cascata di particelle elettricamente cariche e pertanto questa interazione viene osservata come un cono singolo di radiazione Cerenkov.

La tecnica fondamentale di tutti gli esperimenti è quella di compensare la estrema lentezza del decadimento del protone controllando accuratamente una massa di materiale molto grande. Maggiore è la massa, maggiore è il numero di protoni e di neutroni legati e quindi maggiore la probabilità di osservare un decadimento. In questo modo ci si aspetta che sia possibile rivelare il decadimento del protone o del neutrone legato con una vita media piuttosto lunga rispetto al limite attuale di 10^{30} anni. Gli esperimenti differiscono principalmente per la natura e per la quantità di materiale controllato, la natura e la disposizione degli strumenti usati per rivelare un decadimento di un protone nel materiale e le caratteristiche dell'esperimento adottate per eliminare segnali spuri dei raggi cosmici, tra le quali la profondità del sottosuolo alla quale gli esperimenti vengono eseguiti.

Dal momento che deve essere controllata una massa grandissima, gli esperimenti devono fare impiego di qualche materiale relativamente poco costoso, quale l'acqua, il cemento o il ferro. In esperimenti quali quelli della miniera di Soudan, del campo aurifero di Kolan e del tunnel del Monte Bianco, che impiegano ferro o cemento, ci si deve affidare a dispositivi di rivelazione quali i contatori proporzionali o i contatori a scarica, che

possono rivelare direttamente le particelle cariche energetiche che si prevede siano emesse nei decadimenti del protone. Le particelle cariche hanno un percorso limitato nel ferro o nel cemento e quindi i rivelatori devono essere disposti a piccole distanze attraverso il materiale che viene controllato.

Invece, esperimenti quali quelli della miniera di sale di Morton, della miniera di Silver King e della miniera d'oro Homestake, che controllano un materiale trasparente quale l'acqua, possono fare ricorso a una strategia alquanto differente. L'energia liberata nel decadimento del protone è sufficientemente grande perché un elettrone, un positone, un muone o un pione emessi nel processo di decadimento possano avere un'altissima velocità, ovviamente inferiore a quella della luce nel vuoto, ma superiore alla velocità della luce nell'acqua. Quando una particella carica attraversa un mezzo trasparente con una velocità superiore a quella della luce in quel mezzo, si verifica il fenomeno conosciuto come effetto Cerenkov. È simile al *bang* sonoro generato da un aereo che viaggia con velocità superiore a quella del suono in aria, ma l'effetto Cerenkov è un *bang* ottico, nel quale la particella emette un cono di luce anziché di suono. (Lo spettacolare bagliore azzurro della luce Cerenkov è stato notato nei primi esperimenti di radioattività di Marie Curie, ma le sue proprietà furono scoperte per la prima volta in dettaglio negli anni trenta da Pavel A. Cerenkov.) L'angolo tra i raggi della luce Cerenkov e la traiettoria delle particelle cariche dipende dal rapporto tra la velocità delle particelle cariche e la velocità della luce nel mezzo. La maggior parte della luce viene emessa da particelle che viaggiano quasi con la velocità della luce nel vuoto e in questo caso l'angolo ha nell'acqua un valore tipico di circa 42 gradi.

L'osservazione di un cono di luce Cerenkov è un segnale che è stata creata nel mezzo una particella carica veloce. Inoltre, per una data velocità iniziale, lo spessore del cono e la quantità di luce emessa dipendono soltanto dalla distanza che la particella carica percorre prima che la sua velocità scenda sotto la velocità della luce nel mezzo, distanza che a sua volta dipende dalla sua energia iniziale. Perciò, registrando le posizioni nelle quali la luce viene ricevuta e l'intensità della luce, si può dedurre sia l'energia iniziale sia la direzione di ciascuna particella carica. In certe circostanze potrebbe decadere la stessa particella in moto, emettendo altre particelle che potrebbero dare origine a un secondo lampo di luce Cerenkov. Un muone o un antimuone creati in un decadimento di un protone potrebbero decadere in un elettrone che emette radiazione Cerenkov. Un pione carico potrebbe decadere in un muone lento (o in un antimuone lento) che a sua volta potrebbe decadere in un elettrone (o in un positone) veloce che dovrebbe emettere un cono di luce. Un pione neutro potrebbe decadere in due fotoni, ciascuno dei quali potrebbe originare uno sciame di particelle cariche e la conseguente luce Cerenkov. L'impiego della luce Cerenkov per rivelare i decadimenti del protone offre perciò un'alternativa all'osservazione diretta di particelle cariche come metodo per ricostruire il processo di decadimento e verificare che stiamo realmente osservando un decadimento di un protone.

Un vantaggio dell'impiego della luce Cerenkov su altri metodi di rivelazione del decadimento del protone è che la luce può percorrere in acqua distanze maggiori di quelle che possono percorrere le stesse particelle cariche. Quindi, con un dato volume di materiale sotto controllo è necessario un minor numero di rivelatori che non in esperimenti che impiegano materiali opachi quali il ferro o il cemento. Inoltre l'acqua è meno costosa del ferro e del cemento. Del resto, i rivelatori Cerenkov ad acqua sono sensibili soltanto alle particelle cariche che viaggiano con velocità maggiore di quella della luce nell'acqua. Inoltre, la densità relativamente bassa dell'acqua richiede che vengano scavate nel sottosuolo cavità maggiori per contenere una data massa di materiale controllato e l'acqua deve essere mantenuta molto pura perché conservi la sua trasparenza alla luce Cerenkov. Come ho detto, la ricerca sta procedendo alacremente sia per gli esperimenti che impiegano l'acqua come materiale controllato sia per quegli esperimenti che impiegano materiali più densi.

Quali frequenze di decadimenti del protone si potrebbero rivelare in tali esperimenti? Per fare un esempio, l'esperimento che prevede il controllo della massa maggiore è quello della miniera di sale di Morton. Della sua massa controllata di 10 000 tonnellate d'acqua, un involucro esterno di circa 5000 tonnellate sarà impiegato per rivelare le tracce dei raggi cosmici di fondo. La restante massa di circa 5000 tonnellate d'acqua comprende 3×10^{33} protoni (e neutroni legati). Se il protone ha la vita media di circa 10^{31} anni indicata dalle versioni perfezionate dell'analisi eseguita da Georgi, Quinn e da me, dovrebbero riscontrarsi circa 300 decadimenti del protone all'anno. Alcuni anni di osservazione porterebbero a qualche evento di decadimento del protone anche se la vita media fosse lunga 10^{33} anni, ma con frequenze di decadimento così basse gli esperimenti comincerebbero a essere messi in forse da un non eliminabile fondo di eventi spuri dovuti ai neutrini dei raggi cosmici e sarebbero difficili ulteriori miglioramenti.

Che cosa verremmo a sapere di nuovo se venisse scoperto il decadimento del protone? Naturalmente si concluderebbe immediatamente che il numero barionico non si conserva e ciò sarebbe di sostegno all'opinione diffusa che tutte le grandezze conservate hanno un significato dinamico simile a quello della carica elettrica. Inoltre, se il decadimento del protone sarà scoperto nel prossimo futuro, la sua vita media dovrà risultare compresa tra 10^{30} e 10^{33} anni e questo fatto darà una certa credibilità alle ipotesi generali sull'unificazione delle forze forte ed elettrodebole avanzate da Georgi, da Quinn e da me. Esiste però un gran numero di teorie che soddisfano queste ipotesi generali, compresa almeno qualche versione dei modelli $SU(4)^4$, $SU(5)$, $SO(10)$ e di altri sopra citati. Sarà difficile dire quale di queste teorie (ammesso che ve ne sia qualcuna) descriva la fisica delle altissime energie.

Di una sola cosa possiamo esser sicuri. Se si scoprirà il decadimento del protone, verranno dedicate nuove grandi risorse al suo studio e quanto prima ci sarà una seconda generazione di esperimenti il cui obiettivo sarà quello di stabilire non se il protone decade, ma come decade: quali sono le probabilità dei vari modi di decadimento?

In vista di tale sforzo numerosi teorici hanno esplorato i probabili modi di decadimento del protone. (Le considerazioni che seguono si basano su ricerche indipendenti di Wilczek, di Anthony Zee e mie.) È interessante il fatto che ci si possa spingere sufficientemente lontano in questa analisi senza alcuna ipotesi sull'unificazione delle interazioni forte ed elettrodebole. Tutto ciò che si richiede sono i familiari principi di conservazione della carica, del colore e così via e l'ipotesi che le particelle esotiche responsabili del decadimento del protone siano molto pesanti, come esse devono certamente essere per spiegare la lunga vita media del protone. In questo caso, anche se l'emissione e il riassorbimento di queste particelle possono produrre un gran numero di modi diversi di decadimento del protone, l'ipotesi della grande massa delle particel-

LOCALITÀ	DATA DI FUNZIONAMENTO	PROFONDITÀ (METRI D'ACQUA EQUIVALENTI)	MATERIALE CONTROLLATO	RIVELATORI	ISTITUZIONI
MINIERA AURIFERA DI HOMESTAKE, SOUTH DAKOTA	OGGI	4400	150 TONNELLATE D'ACQUA (DA AUMENTARE A 900 TONNELLATE)	4 FOTOMOLTIPLICATORI ACQUA	UNIVERSITÀ DELLA PENNSYLVANIA
CAMPO AURIFERO DI KOLAR, INDIA	OGGI	7600	150 TONNELLATE DI FERRO (LASTRE DA 1/2 POLLICE)	)NTATORI PROPORZIONALI A GAS IA LE LASTRE	TATA INSTITUTE OF FUNDAMENTAL RESEARCH, BOMBAY UNIVERSITÀ DI OSAKA UNIVERSITÀ DI TOKYO
VALLE DI BAKSAN, URSS	OGGI	850	80 TONNELLATE DI SCINTILLATORE LIQUIDO	00 FOTOMOLTIPLICATORI	ISTITUTO PER LA RICERCA NUCLEARE, MOSCA
MINIERA DI SOUDAN, MINNESOTA	OGGI	1800	30 TONNELLATE DI CEMENTO A TACONITE E DI FERRO (DA AUMENTARE A 1000 TONNELLATE)	56 CONTATORI PROPORZIONALI GAS NEL CEMENTO	UNIVERSITÀ DEL MINNESOTA ARGONNE NATIONAL LABORATORY
TUNNEL DEL MONTE BIANCO, TRA FRANCIA E ITALIA	OGGI	4270	30 TONNELLATE DI FERRO E DI SCINTILLATORE LIQUIDO (DA AUMENTARE A 200 TONNELLATE)	)TOMOLTIPLICATORI TUBI A STREAMER	ISTITUTO PER LA RICERCA NUCLEARE, MOSCA UNIVERSITÀ DI TORINO
MINIERA DI SALE DI MORTON, OHIO	PREVISTA ENTRO IL 1981	1670	10 000 TONNELLATE D'ACQUA	00 FOTOMOLTIPLICATORI ACQUA	UNIVERSITÀ DELLA CALIFORNIA A IRVINE UNIVERSITÀ DEL MICHIGAN BROOKHAVEN NATIONAL LABORATORY
MINIERA DI SILVER KING, UTAH	PREVISTA ENTRO IL 1981	1700	1000 TONNELLATE D'ACQUA	0 FOTOMOLTIPLICATORI ACQUA E PARETI SPECULARI	HARVARD UNIVERSITY PURDUE UNIVERSITY UNIVERSITÀ DEL WISCONSIN
TUNNEL DEL MONTE BIANCO, TRA FRANCIA E ITALIA	PREVISTA ENTRO IL 1981	5000	INIZIALMENTE 150 TONNELLATE DI FERRO (LASTRE DA UN CENTIMETRO)	JBI A STREAMER RA LE LASTRE	LABORATORI NAZIONALI DI FRASCATI UNIVERSITÀ DI MILANO UNIVERSITÀ DI ROMA UNIVERSITÀ DI TORINO
TUNNEL DEL GRAN SASSO, ITALIA	IN PROGETTO	4000	10 000 TONNELLATE DI FERRO (LASTRE DA TRE MILLIMETRI)	AMERE A FLASH INNESCATE A TUBI A STREAMER DA CONTATORI LASTRE RESISTIVE	LABORATORI NAZIONALI DI FRASCATI UNIVERSITÀ DI MILANO UNIVERSITÀ DI TORINO UNIVERSITÀ DI ROMA
GIAPPONE	IN ESAME	2700	3400 TONNELLATE D'ACQUA	56 FOTOMOLTIPLICATORI A 20 POLLICI	UNIVERSITÀ DI TOKYO LABORATORIO NAZIONALE GIAPPONESE PER LA FISICA DELLE ALTE ENERGIE UNIVERSITÀ DI TSUKUBA
TUNNEL DEL FRÉJUS, TRA FRANCIA E ITALIA	IN ESAME	4500	1500 TONNELLATE DI FERRO (LASTRE DA TRE A QUATTRO MILLIMETRI DI SPESSORE)	AMERE A FLASH INNESCATE A TUBI A STREAMER	ÉCOLE POLYTECHNIQUE UNIVERSITÀ DI PARIGI A ORSAY ÉCOLE DES MINES A SACLAY
GIAPPONE	IN ESAME	2700	600 TONNELLATE DI FERRO	AMERE A FLASH ON TUBI AL NEO A CONTATORI GEIGER	UNIVERSITÀ DI OSAKA UNIVERSITÀ DI TOKYO LABORATORIO NAZIONALE GIAPPONESE PER LA FISICA DELLE ALTE ENERGIE
MINIERA DI SALE DI ARTYEMOVSK, URSS	IN ESAME	600	100 TONNELLATE DI SCINTILLATORE LIQUIDO	8 FOTOMOLTIPLICATORI	ISTITUTO PER LA RICERCA NUCLEARE, MOSCA

La tabella degli esperimenti sul decadimento del protone comprende cinque rivelatori oggi in funzione, tre rivelatori che è previsto entrino in funzione quest'anno e cinque rivelatori in fase di studio. Gli esperimenti differiscono nel genere e nella quantità di materiale controllato, e vengono eseguiti a profondità diverse nel sottosuolo. La profondità è espressa in termini della profondità d'acqua equivalente che fornirebbe lo stesso schermaggio dai raggi cosmici.

Le ricerche in Italia

Nel garage 17 del tunnel del Monte Bianco una collaborazione tra ricercatori dei Laboratori nazionali di Frascati, della sezione di Milano dell'Istituto nazionale di fisica nucleare (INFN) e dell'Istituto di cosmogeofisica di Torino del CNR prepara, con il supporto del CERN, l'installazione di un apparato cubico (43 metri cubi) di 150 tonnellate, costituito da 134 piani di ferro spessi 1 centimetro, intervallato con strati di tubi a *streamer*, una tecnica di rivelazione sviluppata a Frascati. Il ferro costituisce sia il campione di nucleoni osservati (circa 10^{32}), sia il materiale di interazione delle particelle prodotte nel decadimento per permetterne l'identificazione.

Le particelle cariche ionizzano il gas di riempimento dei tubi, provocando scariche filiformi *(streamer)* nell'intenso campo elettrico prodotto mediante un filo metallico teso all'interno dei tubi. Queste scariche inducono segnali elettrici su un reticolo di circa 100 000 strisce metalliche, collegate a circuiti elettronici per la registrazione e presentazione visiva dell'evento mediante calcolatore. Il disegno qui sotto mostra come potrebbe apparire in questo apparato il decadimento di un protone in un positone e in un mesone π^0. Tale previsione risulta da misure di calibrazione effettuate su una frazione già realizzata dell'apparato, esposta a vari fasci di particelle al Protosincrotrone del CERN.

Nel tunnel del Gran Sasso, sull'autostrada Teramo-L'Aquila, sorgerà, su iniziativa dell'INFN, un laboratorio sotterraneo di vaste proporzioni (100 × 25 × 25 metri cubi), che porrà l'Italia in una posizione di privilegio nel campo di tutta quella sperimentazione scientifica che necessita di una forte attenuazione del flusso di raggi cosmici, operata in questo caso da uno spessore di roccia equivalente a circa 4000 metri di acqua. Esperimenti sulla stabilità del protone innanzi tutto, ma anche in altri campi della fisica delle particelle, dell'astrofisica ecc. Questo laboratorio godrà anche dell'essenziale vantaggio di essere ben collegato ai Laboratori nazionali di Frascati, nascendo di fatto a essi strettamente associato e da esso dipendente per le strutture necessarie alla realizzazione degli imponenti apparati sperimentali ora allo studio.

Hanno manifestato interesse per questo tipo di sperimentazione numerosi fisici delle università italiane e dell'INFN (Frascati, Milano, Roma, Torino), oltre che stranieri (Canada, Cina, Francia, Giappone).

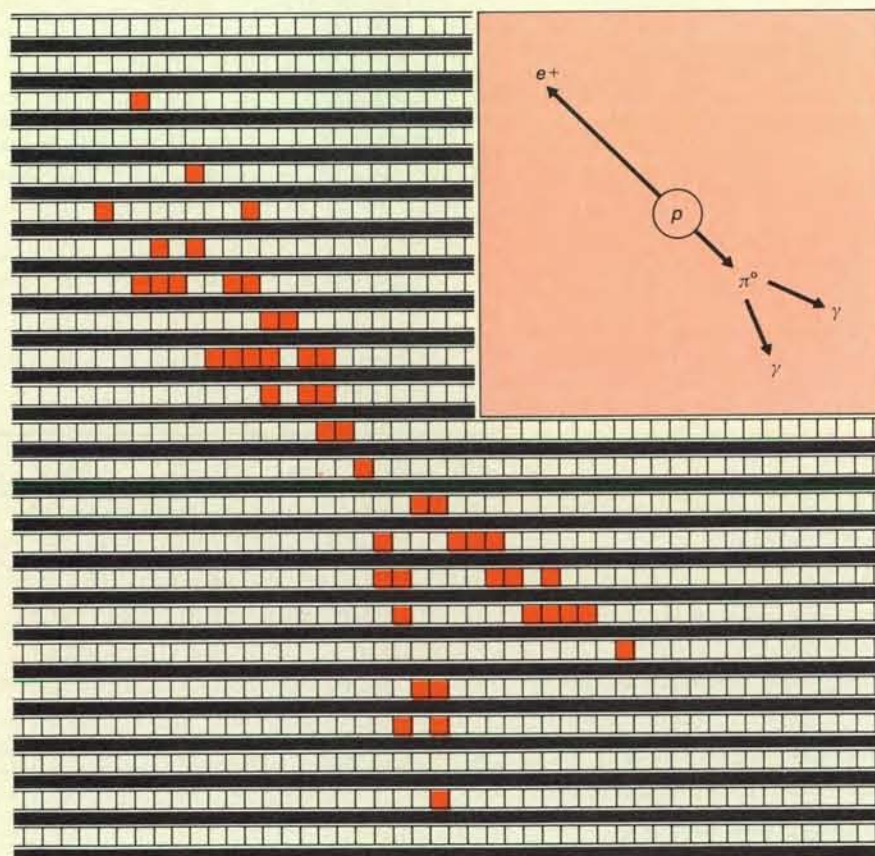
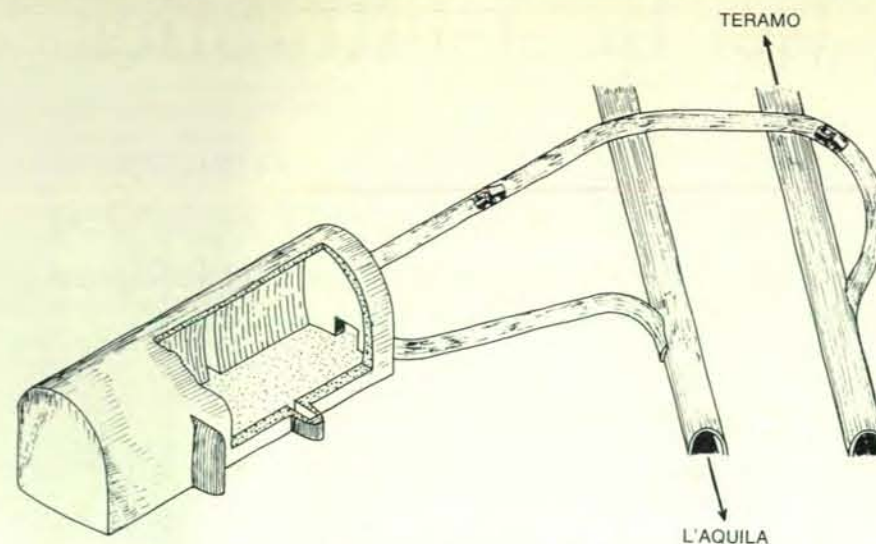
La tendenza attuale mira alla realizzazione di apparati del tipo visualizzante, con masse nel campo delle migliaia di tonnellate.

Infatti presso alcuni Istituti di ricerca italiani è attualmente allo studio il progetto di un gigantesco rivelatore di tracce a struttura modulare, la cui massa potrà estendersi nella regione delle 10 000 tonnellate, adatto a essere installato nel futuro laboratorio sotterraneo del Gran Sasso per «vedere» il decadimento del protone e per studiare altri fenomeni di grande interesse fisico e astrofisico («oscillazioni» dei neutrini prodotti dai raggi cosmici nell'atmosfera terrestre, sorgenti galattiche ed extra-galattiche di neutrini di alta energia ecc.).

Nella figura in basso della pagina a fronte sono mostrati due esempi di decadimento del protone così come apparirebbero in uno dei moduli del rivelatore, costituiti da una successione di «camere a flash» di circa 60 metri quadrati di area utile, alternate con sottili lastre di ferro (3 millimetri di spessore) e, per ogni 10 di tali alternanze, con piani di contatori («calorimetro a flash»).

Le camere sono costituite da grandi lastre di polipropilene estruso formate da un elevato numero di alveoli tubolari lunghi 8 metri, a sezione quadrata di 4 millimetri di lato, nei quali fluisce una miscela di gas nobili. Vi sono più di un milione di alveoli per ogni modulo da 1000 tonnellate.

Le particelle cariche emesse nel decadimento del protone producono dei segnali elettrici nei contatori da esse attraversati. Questi segnali vengono elaborati da un opportuno sistema elettronico per ottenere, immediatamente dopo che il protone è decaduto, un impulso elettrico di alta tensione che viene applicato alle camere. Le tracce delle particelle cariche emesse nel decadimento appaiono allora come una successione di piccoli lampi di luce *(flash)* emessi dagli alveoli attraversati dalle particelle stesse. Infatti ognuno di tali alveoli diventa sede di una scarica elettrica luminosa che, innescata dagli elettroni di ionizzazione per effetto dell'intenso campo elettrico impulsivo, si propaga per tutta la lunghezza dell'alveolo.

Nell'evento disegnato in alto nella figura citata un protone decade in un antimuone (μ^+) e in un mesone K^0. L'antimuone viene emesso con una energia cinetica di 246 milioni di elettronvolt e percorre circa 18 centimetri di ferro prima di arrestarsi dopo aver perso tutta la sua energia cinetica negli urti con gli atomi del mezzo. Il mesone K^0 (di cui non si vede naturalmente la traccia perché è elettricamente neutro) decade in due pioni carichi (π^+ e π^-) dopo alcuni decimi di miliardesimo di secondo durante i quali esso percorre qualche centimetro nel rivelatore.

L'apparato con tubi a **streamer.**

Il laboratorio del Gran Sasso.

Nell'evento in basso, un protone decade in un positone e in un pione neutro, il quale a sua volta decade in un tempo brevissimo (non misurabile nel rivelatore) in due fotoni di alta energia (γ_1 e γ_2). Sia il positone sia i due raggi γ danno luogo, nell'attraversare le lastrine di ferro del rivelatore, allo sviluppo di «sciami elettrofotonici» che hanno l'aspetto mostrato in figura. Le energie del positone e del mesone si possono determinare contando i numeri degli alveoli accesi nei tre sciami.

Le camere a flash in materiale plastico estruso utilizzate in questo progetto e nell'analogo progetto francese sviluppato con il concorso di alcuni fisici italiani (un rivelatore di 1500 tonnellate da installare nel tunnel del Fréjus) sono una versione nuova (sviluppata all'Università di Roma nella tesi di laurea di Lorenzo Federici nel 1973) di una tecnica ideata nel 1955 all'Università di Pisa da Marcello Conversi e da Adriano Gozzini, tecnica poi vastamente impiegata in ricerche sui raggi cosmici e che è stata all'origine di una nuova famiglia di rivelatori di tracce (camere a scintilla, camere a *streamer* ecc.).

I contatori adoperati per sensibilizzare le camere a flash saranno probabilmente o i tubi a streamer limitato, sviluppati a Frascati dal gruppo di Enzo Iarocci negli anni 1977-1979, oppure i «contatori a lastre resistive», di un nuovo tipo, sviluppati recentemente da Rinaldo Santonico all'Istituto fisico dell'Università di Roma. Con i contatori a lastre resistive è possibile raggiungere risoluzioni temporali dell'ordine del miliardesimo di secondo, adatte a misure di «tempi di volo».

(Su informazioni dei Laboratori nazionali di Frascati INFN e dell'Istituto fisico dell'Università di Roma).

PIANI DI CONTATORI PER IL COMANDO DELLE CAMERE A FLASH

Due possibili decadimenti del protone in un «calorimetro a flash».

le esotiche è sfavorevole in forma più decisa ai modi di decadimento più complessi che non a quelli più semplici. A meno che non si verifichi qualche particolare circostanza, i modi di decadimento dominanti sono in generale quelli nei quali il protone o il neutrone legato decadono in un positone, un antimuone o un antineutrino, oltre a un certo numero di mesoni, piuttosto che quelli in cui il decadimento avviene in un elettrone, un muone o un neutrino, oltre a un certo numero di mesoni. Si può andare oltre e fare previsioni sui rapporti dei modi di decadimento. Per esempio, un neutrone decade due volte più rapidamente di un protone in un positone e un solo mesone pi o mesone rho. Il protone decade in un positone e in mesoni più rapidamente di quanto un neutrone decada in un antineutrino e in mesoni. Il neutrone decade più rapidamente in un positone e mesoni di quanto il protone decada in un antineutrino e mesoni.

Non si può esser certi che tali previsioni verranno confermate dagli esperimenti. Se non sarà così, devono esistere particelle esotiche nettamente più leggere di 10^{14} masse protoniche per produrre modi di decadimento più complessi. Per esempio, il decadimento di un protone o di un neutrone legato in un elettrone e mesoni anziché in un positone e mesoni si potrebbe produrre a una frequenza osservabile se esistessero particelle esotiche non più pesanti di circa 10^{10} masse protoniche. Un decadimento in tre neutrini o tre elettroni (o qualche altra combinazione di tre leptoni) si potrebbe osservare se esistessero particelle esotiche non più pesanti di circa 10^4 masse protoniche. Tuttavia, tali particelle esotiche relativamente leggere dovrebbero godere di particolari proprietà per evitare di produrre il decadimento «ordinario» del protone (in un mesone e un positone o un antineutrino) a frequenza troppo alta.

La verifica di queste previsioni, cioè la scoperta che il protone decade in mesoni e un positone o un antineutrino e con i sopra citati rapporti di decadimento, servirebbe a confermare che il decadimento del protone è davvero dovuto a particelle esotiche con masse maggiori di circa 10^{10} masse protoniche, ma non indicherebbe come privilegiata alcuna teoria fondamentale. A tale scopo sarebbe necessario esplorare dettagli più fini del processo di decadimento. (Per esempio, una determinazione della direzione nella quale ruotano un positone o un antimuone quando vengono prodotti nel decadimento di un protone si potrebbe impiegare per diagnosticare lo spin delle particelle esotiche superpesanti la cui emissione e riassorbimento hanno prodotto il decadimento.) Se si scoprirà il decadimento del protone, esso segnerà un trionfo dell'ingegno sperimentale e costituirà una chiave senza precedenti per la fisica delle altissime energie, ma metterà davanti ai fisici sperimentali e teorici molti nuovi compiti che dovranno essere assolti se si vorrà comprendere il meccanismo del decadimento del protone.

Gli antibiotici beta-lattamici

Le penicilline e le cefalosporine sono caratterizzate chimicamente dall'anello beta-lattamico. La loro struttura molecolare è stata modificata per aumentarne l'efficacia e combattere i batteri resistenti

di E. P. Abraham

Sono passati quasi 40 anni da quando la penicillina è stata introdotta nella pratica medica. Non molto tempo prima il trattamento con sulfonammidi di pazienti con febbri puerperali e altre pericolose malattie causate dagli streptococchi aveva dimostrato per la prima volta che era possibile curare efficacemente le infezioni batteriche generalizzate per mezzo della somministrazione di piccole molecole. Molti di coloro che praticano la medicina oggi non hanno avuto diretta esperienza del tempo in cui non era possibile fare quasi nulla per aiutare i pazienti affetti da una seria infezione batterica, quando l'endocardite batterica era quasi sempre fatale, quando la meningite da meningococchi provocava infermità gravissime in quei pochi che sopravvivevano, e quando la polmonite pneumococcica era nota come «l'amica dei vecchi», poiché era una causa di morte indolore. La penicillina, con la sua elevata attività contro molti batteri patogeni e la sua tossicità quasi trascurabile per l'uomo, sembrò possedere proprietà quasi miracolose e segnò l'inizio di una nuova era nel trattamento chemioterapico delle malattie.

In seguito al successo della penicillina, una estesa ricerca internazionale permise la scoperta di una grande quantità di altre sostanze antibatteriche naturali. Molte di queste erano altamente tossiche e furono presto abbandonate. Altre, come la bacitracina, si dimostrarono di utilità marginale nella medicina umana. Quelle clinicamente utili comprendono la streptomicina (e altri amminoglicosidi come la gentamicina), le eritromicine e la famiglia delle tetracicline. Le prime penicilline furono sostituite da una varietà di penicilline modificate che si dimostrarono attive contro i ceppi batterici che erano diventati resistenti alla penicillina. Lo studio del meccanismo della resistenza agli antibiotici e della loro attività è diventato un ramo molto attivo della biologia cellulare e della genetica. La famiglia delle penicilline è stata anche ampliata con l'inclusione delle cefalosporine, composti chimici correlati alle penicilline e scoperti all'Università di Oxford; attualmente le cefalosporine assorbono all'incirca il 15-20 per cento di tutte le spese sostenute per prodotti farmaceutici negli ospedali degli Stati Uniti.

L'evento critico che aprì l'era degli antibiotici in medicina fu la decisione di Howard W. Florey ed Ernst B. Chain nel 1938 di iniziare uno studio sistematico, presso la Sir William Dunn School of Pathology a Oxford, delle sostanze antibatteriche prodotte dai microrganismi. Florey e Chain erano motivati principalmente da interessi scientifici e non dalla speranza che il progetto avrebbe portato a sostanze di valore clinico. Fortunatamente la penicillina fu una delle tre sostanze scelte per essere studiate per prime.

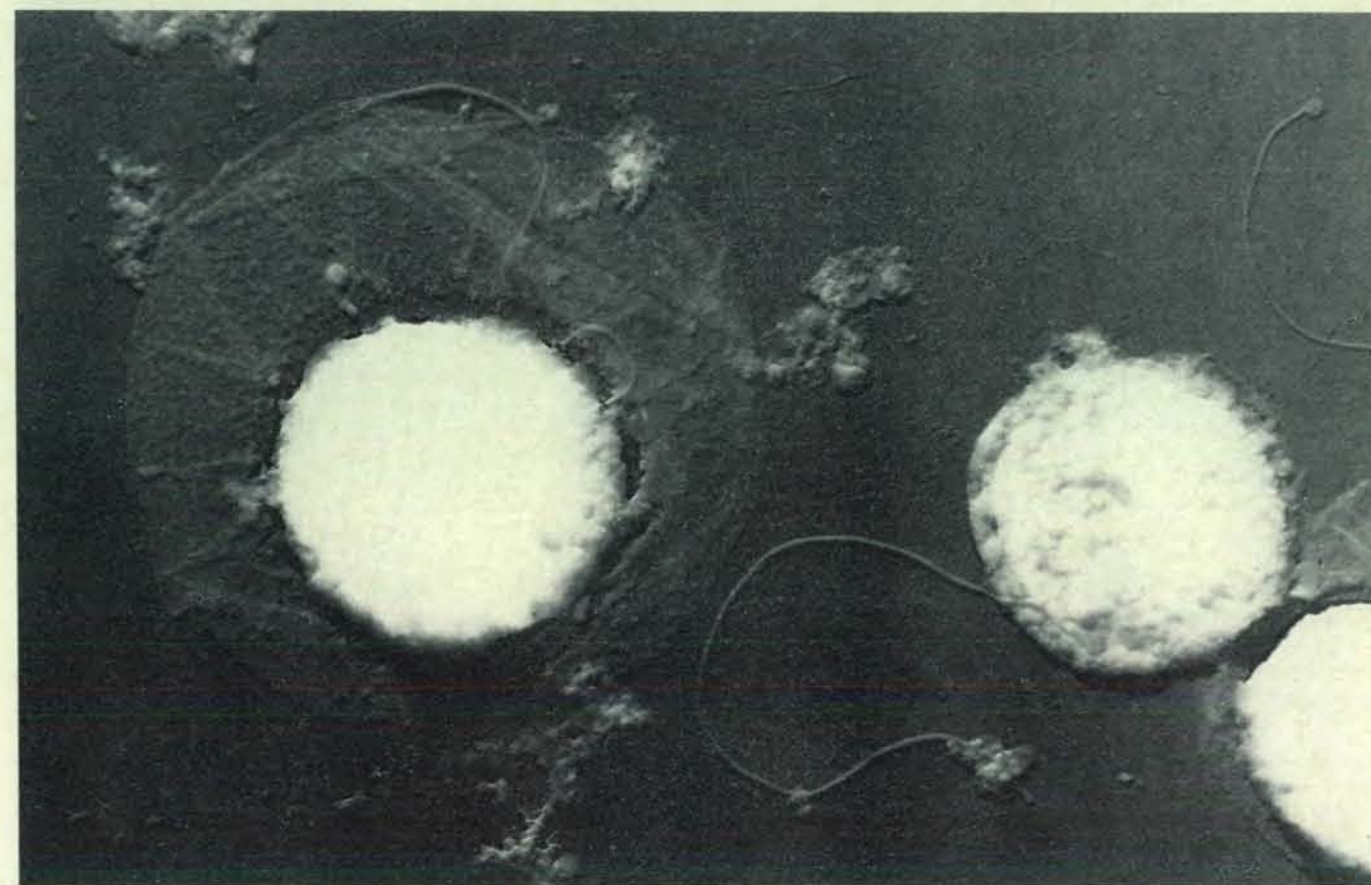
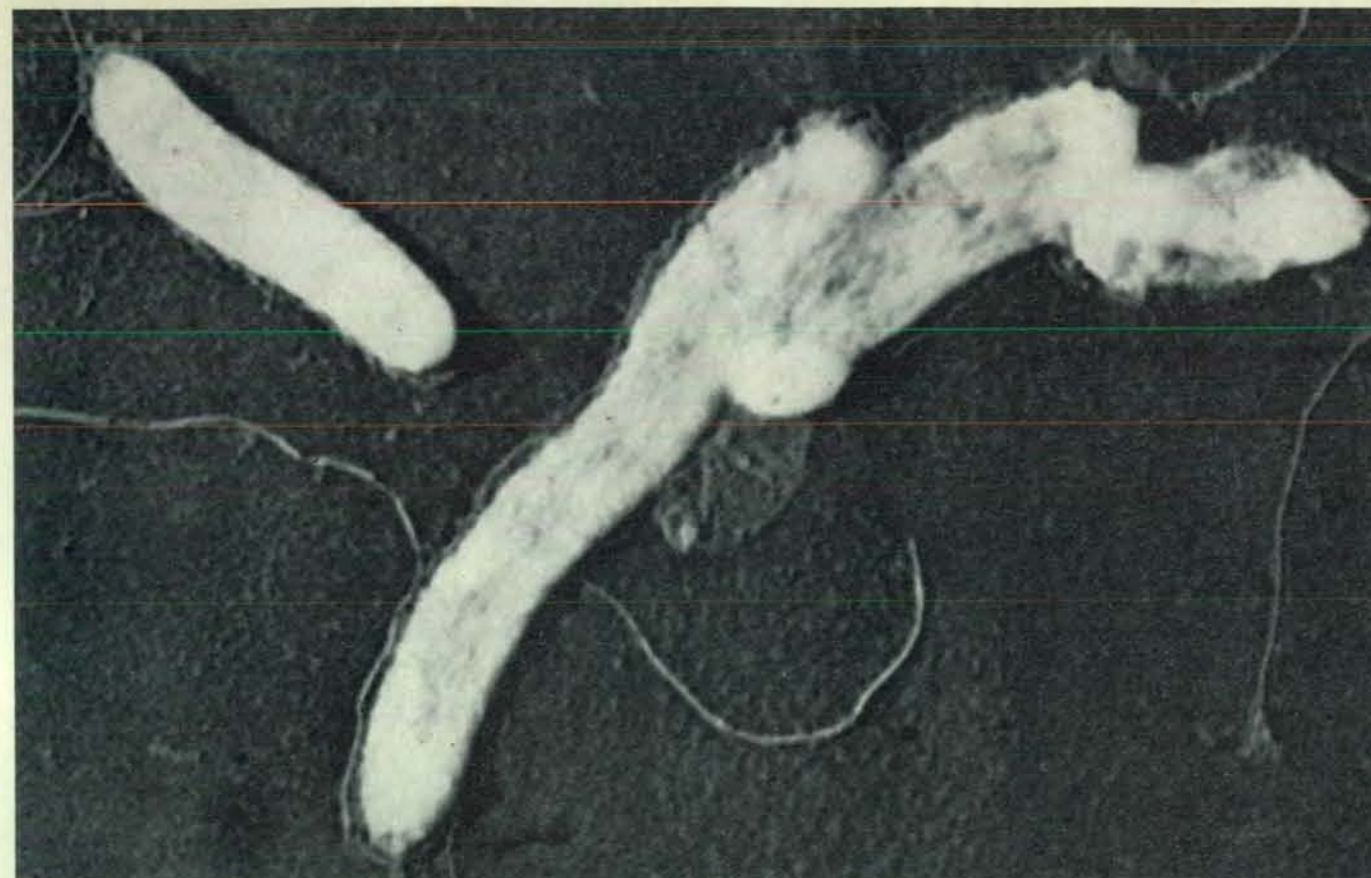
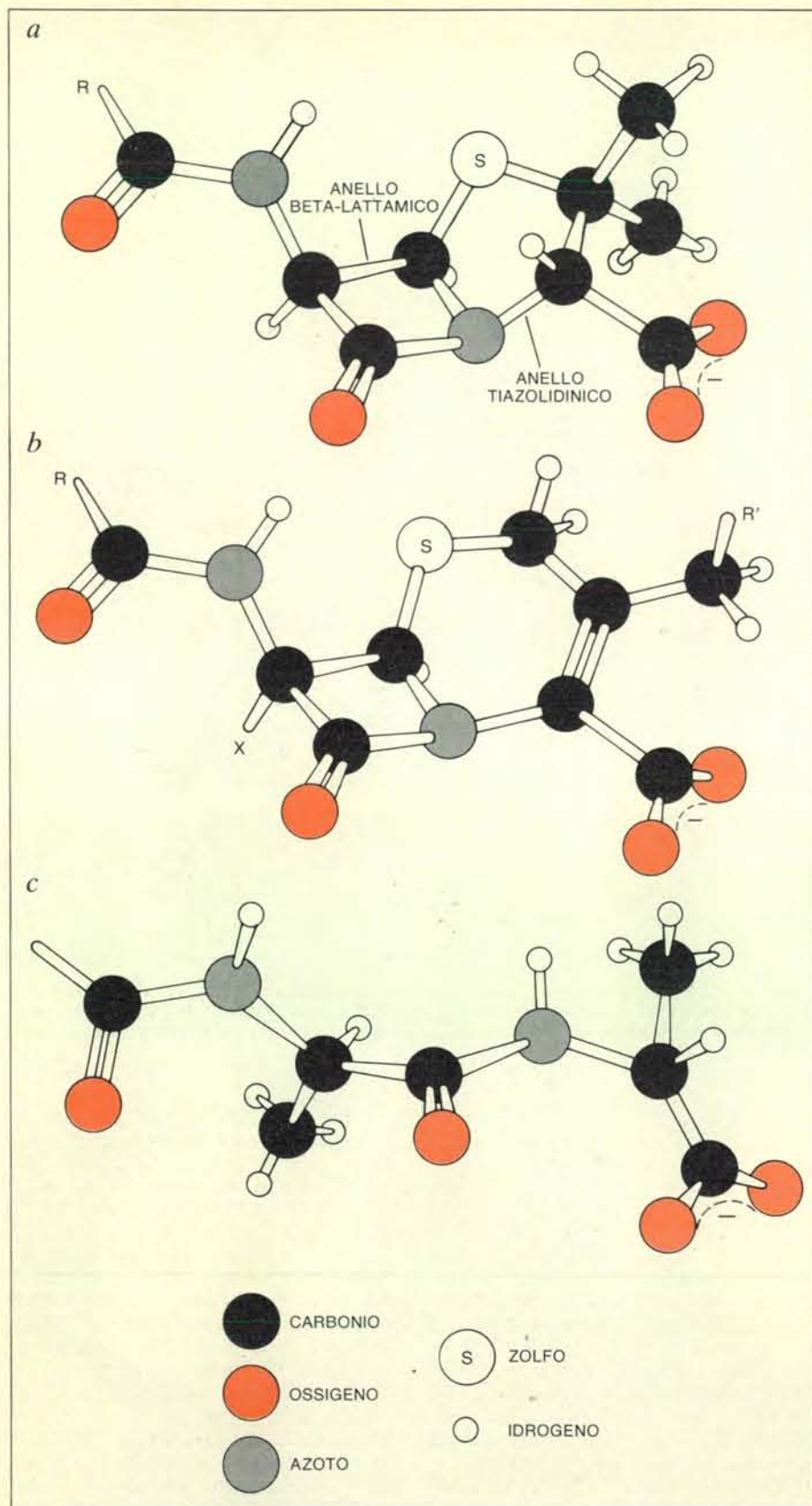
La penicillina in sé non era nuova. Era il nome dato da Alexander Fleming nel 1929 a un «filtrato di brodo di muffa» con attività antibatterica che egli aveva ottenuto dopo la crescita di un fungo, il *Penicillium notatum*, in un brodo nutritivo al St. Mary's Hospital di Londra. Fleming stava studiando delle varianti degli stafilococchi, i batteri che possono produrre foruncoli, ascessi e talvolta infezioni generalizzate e fatali. Il fungo era capitato per caso su una piastra di agar nutritivo su cui Fleming aveva seminato i batteri e che poi aveva messo da parte prima delle vacanze estive. Un periodo di tempo freddo permise al fungo di crescere e secernere la penicillina prima che la crescita dei batteri, che avevano bisogno di una temperatura più alta, fosse terminata. Al suo ritorno al laboratorio Fleming osservò che le colonie dei batteri in vicinanza della colonia di *Penicillium* erano parzialmente lisate, cioè dissolte. La lisi non sarebbe avvenuta se il tempo fosse stato più caldo, permettendo agli stafilococchi di crescere per primi: la penicillina provoca la lisi dei batteri in crescita, ma non di quelli che hanno terminato di crescere.

Fleming non era stato il primo a osservare che un fungo del genere *Penicillium* impediva la crescita dei batteri nelle sue vicinanze. Questo fenomeno era stato notato da John Burdon Sanderson nel 1870, da Joseph Lister un anno dopo, da William Roberts nel 1874, da John Tyndall nel 1876 e da uno studioso francese, Ernest Duchesne, nel 1896. Non si saprà mai se qualcuno di questi ricercatori avesse osservato gli effetti della penicillina, poiché le specie di *Penicillium* secernono molte altre sostanze antibatteriche.

In ogni caso Fleming fu un acuto osservatore di questo fenomeno inatteso; ne fu interessato e lo studiò più a fondo. Dimostrò che la penicillina era attiva contro numerose specie di batteri Gram-positivi, come gli stafilococchi, gli streptococchi e i pneumococchi, ma non contro una varietà di bacilli Gram-negativi patogeni come *Salmonella typhi, Escherichia coli* e *Pseudomonas aeruginosa*. I batteri Gram-positivi e Gram-negativi si distinguono per la diversa risposta alla procedura di colorazione ideata da Hans Christian Joachim Gram nel 1884, che permette di evidenziare importanti differenze nella parete cellulare degli organismi.

Fleming dimostrò inoltre che il suo brodo contenente penicillina non era più tossico di un normale brodo sia se messo a contatto con i globuli bianchi del sangue, sia se iniettato in un coniglio. Tuttavia non si rese conto della importanza della sua scoperta. Sebbene avesse utilizzato la penicillina in numerose occasioni come antisettico locale, nel 1940 scrisse che non sembrava valesse la pena di produrla. Il suo interesse principale per la sostanza era dovuto alla sua azione antibatterica selettiva, che permetteva l'isolamento di certi microrganismi da colture miste. Chiaramente non gli venne mai in mente che la penicillina avrebbe potuto essere introdotta nel sangue per curare le infezioni. Per questo motivo per più di dieci anni dopo la sua scoperta la penicillina rimase poco più che una curiosità.

La situazione cambiò drasticamente quando Florey, Chain e un piccolo gruppo di loro colleghi a Oxford riuscirono a purificare parzialmente la penicillina e dimostrarono, prima nei topi e poi negli esseri umani, che essa possedeva notevoli proprietà terapeutiche. Nel 1941 una prova clinica su piccola scala, limitata dalla scarsità di penicillina (che era talmente preziosa che veniva recuperata dall'urina dei pazienti), fornì il primo di una serie di risultati che Florey descrisse nel 1949 come «in verità così gratificanti da essere a volte quasi incredibili».

Fu questa dimostrazione del valore clinico della penicillina che stimolò i governi

L'effetto della penicillina è mostrato in queste fotografie al microscopio elettronico, eseguite da Milton R. J. Salton della New York University School of Medicine. La microfotografia in alto mostra batteri normali della specie *Vibrio metchnikovi*. La microfotografia in basso mostra gli «sferoplasti» che si generano quando i batteri vengono fatti crescere in presenza di penicillina. Vicino allo sferoplasto a sinistra è visibile la parete cellulare indebolita e staccata del batterio. La penicillina esercita il suo effetto interferendo con la sintesi della parete cellulare durante la crescita dei batteri. Gli sferoplasti sono protetti dalla disintegrazione dalla fragile membrana citoplasmatica, appena visibile in fotografia.

La struttura fondamentale delle penicilline e delle cefalosporine è basata sull'anello beta-lattamico. Nella penicillina *(a)* l'anello beta-lattamico a quattro membri è fuso all'anello tiazolidinico a cinque membri, che consiste di un atomo di zolfo, un atomo di azoto e tre atomi di carbonio. La catena laterale *R* è variabile; nell'illustrazione della pagina a fronte sono raffigurate numerose catene laterali della penicillina. Nelle cefalosporine *(b)*, *R* e *R'* sono entrambi variabili e *X* può essere un atomo di idrogeno o un gruppo metossilico ($-OCH_3$). Si può comprendere come gli antibiotici beta-lattamici esercitano il loro effetto sulla parete cellulare di un batterio in crescita, esaminando la struttura *(c)* di una catena D-alanil-D-alanina di peptidoglicano in crescita di una parete batterica. Una sua conformazione è simile a quella di parte della molecola di penicillina.

della Gran Bretagna e degli Stati Uniti a sostenere il notevole impegno delle società farmaceutiche per la produzione di penicillina in quantità sufficiente per curare i feriti nella seconda guerra mondiale. Il rapido progresso che ne seguì fu dovuto in gran parte al contributo americano: l'introduzione di un terreno di coltura migliorata, la fermentazione sommersa (invece che superficiale) e ceppi mutanti di *Penicillium chrysogenum* a resa più elevata. Queste tre innovazioni risolsero quello che a prima vista era sembrato un problema veramente formidabile. Oggi la resa di penicillina per litro di coltura è molte migliaia di volte superiore alla quantità ottenibile negli anni quaranta a Oxford.

Gli anni della guerra videro anche un intenso sforzo anglo-americano, sostenuto da entrambi i governi, per produrre la penicillina per sintesi chimica. Il successo dipendeva in primo luogo dalla risoluzione della struttura della molecola. Nel 1943 Chain e io avevamo proposto una struttura, fino ad allora ignota, in cui un anello beta-lattamico a quattro membri era fuso a un anello tiazolidinico a cinque membri costituito da un atomo di zolfo, uno di azoto e tre di carbonio. La configurazione beta-lattamica fu successivamente sostenuta con vigore da Robert B. Woodward della Harvard University, ma non fu generalmente accettata fino al 1945, quando una analisi cristallografica ai raggi X, eseguita da Dorothy C. Hodgkin e Barbara W. Low a Oxford, non lasciò più alcun dubbio sulle posizioni relative degli atomi nella molecola.

Nel frattempo i tentativi per sintetizzare la molecola piuttosto instabile della penicillina erano falliti. Fu solo nel 1958 che John C. Sheehan e i suoi colleghi al Massachusetts Institute of Technology realizzarono una sintesi coronata da successo, con l'aiuto di un nuovo reagente che permise loro di chiudere l'anello beta-lattamico in condizioni chimiche non estreme. Malgrado questo successo la sintesi chimica totale non è mai diventata economicamente competitiva con la fermentazione nella produzione di penicillina. La semisintesi a partire da un prodotto di fermentazione, tuttavia, ha portato a nuovi composti di grande valore medico.

Già all'inizio degli anni cinquanta l'avvento della penicillina poteva essere considerato una svolta nella storia delle scienze mediche. L'attività altamente selettiva di questa sostanza contro le cellule batteriche le conferì un'enorme importanza, che mantiene ancora oggi. Ciononostante, era chiaro che la penicillina disponibile allora aveva le sue limitazioni: alcuni batteri erano resistenti a essa. Già nel 1940 Chain e io avevamo trovato un enzima in *E. coli* capace di distruggere la penicillina di Oxford. Suggerimmo che l'enzima, da noi chiamato penicillinasi, fosse implicato nella resistenza dei batteri che lo producevano.

Ancor prima che la struttura caratteristica ad anello della penicillina fosse stata risolta era noto che la penicillina prodotta a Oxford differiva da quella prodotta negli Stati Uniti. Queste due penicilline, chiamate rispettivamente penicillina *F* e penicillina *G*, avevano lo stesso nucleo molecolare, ma catene laterali differenti. La penicillina *G*, nota anche come benzilpenicillina, fu la prima forma largamente usata in medicina e ha conservato la sua importanza fino a oggi. Veniva prodotta negli Stati Uniti perché il *corn-steep* (un sottoprodotto della macerazione del mais) che veniva aggiunto al terreno di coltura conteneva l'acido fenilacetico ($C_6H_5CH_2COOH$), da cui derivava la sua catena laterale.

Più tardi molte penicilline con catene laterali chimicamente correlate furono ottenute fermentando il *Penicillium chrysogenum* con l'aggiunta di precursori appropriati al terreno di coltura. Nessuna di queste penicilline era nettamente superiore alla benzilpenicillina nel suo spettro di attività. Nel 1954, tuttavia, si scoprì che la fenossimetilpenicillina, a differenza della benzilpenicillina, era stabile nell'ambiente acido dello stomaco ed era clinicamente efficace anche somministrata per via orale.

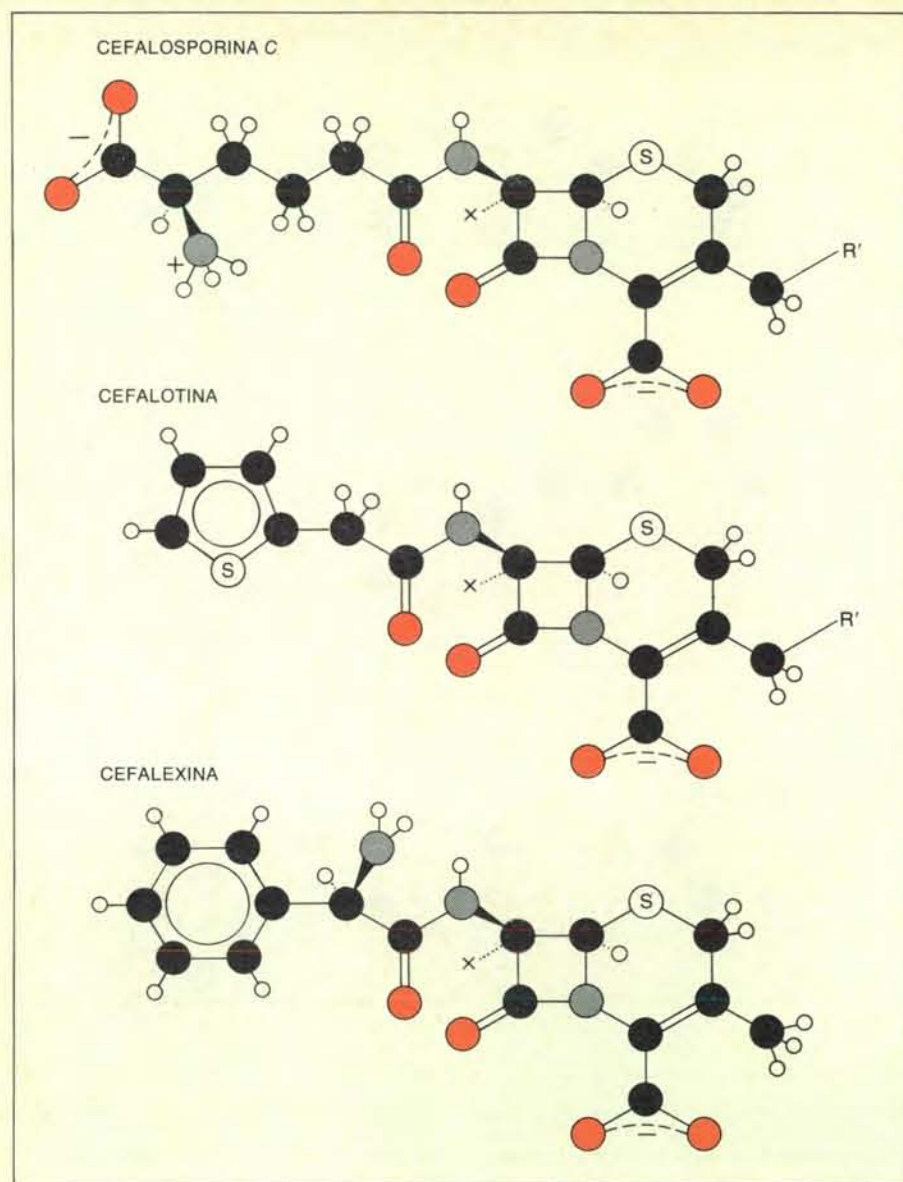
A quel tempo non sembrava possibile ottenere un composto che fosse simile alla penicillina, ma non presentasse le limitazioni della benzilpenicillina. Al contrario, si avvertivano i primi segnali che il valore medico della penicillina sarebbe andato diminuendo poiché gli stafilococchi resistenti alla benzilpenicillina, produttori di penicillinasi, stavano diventando comuni negli ospedali. Fortunatamente numerosi imprevedibili sviluppi portarono alla scoperta di nuove penicilline, della famiglia chimicamente imparentata delle cefalosporine e di altre sostanze contenenti l'anello beta-lattamico che possiedono la maggior parte delle proprietà desiderabili della benzilpenicillina e che sono attive contro molti batteri resistenti alla benzilpenicillina. Questi composti, insieme alle prime penicilline, sono oggi noti collettivamente come antibiotici beta-lattamici. Allo stesso tempo la ricerca ha largamente chiarito il meccanismo di azione degli antibiotici beta-lattamici, le ragioni per cui i diversi membri della famiglia hanno attività differenti, i fattori che possono conferire la resistenza ai batteri e i modi in cui la resistenza può diffondersi.

In seguito ad alcune osservazioni che indicavano che la penicillina uccideva solo i batteri in crescita, Joshua Lederberg, che allora lavorava presso l'Università del Wisconsin, dimostrò che le cellule a forma di bastoncino di *E. coli* che crescevano in un mezzo ipertonico (cioè con una pressione osmotica insolitamente alta) si trasformavano in cellule sferiche in seguito all'aggiunta di penicillina al mezzo di coltura. Questi «sferoplasti» possedevano una fragile membrana citoplasmatica, ma erano privi di gran parte della parete cellulare esterna. Per questo motivo essi scoppiavano quando la pressione osmotica del terreno di crescita veniva ridotta per diluizione. Lederberg dedusse perciò che la penicillina interferiva con la sintesi della parete cellulare dei batteri.

In questo disegno sono rappresentate le catene laterali di tre penicilline. La penicillina *F* è la 2-pentenilpenicillina, la penicillina *G* è la benzilpenicillina e quella *V* è la prima penicillina somministrabile per via orale. Le catene qui raffigurate sono le catene laterali *R* della molecola base.

La comprensione parziale della natura di questa interferenza e delle sue conseguenze è stata accompagnata da un aumento nella conoscenza della struttura molecolare della parete cellulare dei batteri, che è dovuta a Milton R. J. Salton della New York University School of Medicine, a Jack L. Strominger di Harvard, a James Baddiley della Università di Newcastle upon Tyne e a molti altri. Nei batteri Gram-positivi e Gram-negativi la membrana citoplasmatica è coperta da uno strato di peptidoglicano, noto anche come mucopeptide o mureina. Il peptidoglicano forma una struttura tridimensionale il cui scheletro è costituito di unità alternate in due amminozuccheri, la N-acetilglucosammina (NAG) e l'acido N-acetilmuramico (NAM).

Le unità di NAM, che si trovano solo nelle pareti dei batteri e dei microrganismi strettamente correlati, sono collegate a catene peptidiche (corte catene di unità amminoacidiche) in cui sono presenti sia gli amminoacidi D (destro, cioè destrogiri) sia quelli L (levo, cioè levogiri). Molte di queste catene peptidiche a loro volta sono legate tra di loro a formare un reticolo. La struttura che ne risulta conferisce alla parete cellulare la rigidità e alla cellula la sua forma. Lo strato di peptidoglicano dei batteri Gram-negativi, tuttavia, è più sottile di quello dei batteri Gram-positivi e ha meno legami trasversali. Esso, inoltre, è circondato da una membrana esterna che contiene fosfolipidi, lipopolisaccaridi e proteine, e possiede un ambiente interno idrofobo (cioè idrorepellente).

Gli studi di Strominger, di James T. Park della Tufts University School of Medicine e di J.-M. Ghuysen dell'Università di Liegi hanno dimostrato che le membrane citoplasmatiche dei batteri e dei batteri miceliali (cioè simili a funghi) del gruppo degli streptomiceti contengono enzimi che vengono inattivati dalla penicillina e da altri antibiotici beta-lattamici. Gli enzimi vulnerabili sono transpeptidasi (che legano una catena peptidica del peptidoglicano in crescita a un'altra sostituendola a una D-alanina terminale) oppure D-carbossipeptidasi (che rimuovono una D-alanina terminale per idrolisi senza provocare la formazione di legami trasversali) oppure enzimi che possiedo-

In questa pagina sono raffigurate le strutture delle catene laterali di tre cefalosporine. La cefalosporina *C* è stata la prima scoperta della famiglia di penicilline derivate dal fungo *Cephalosporium*.

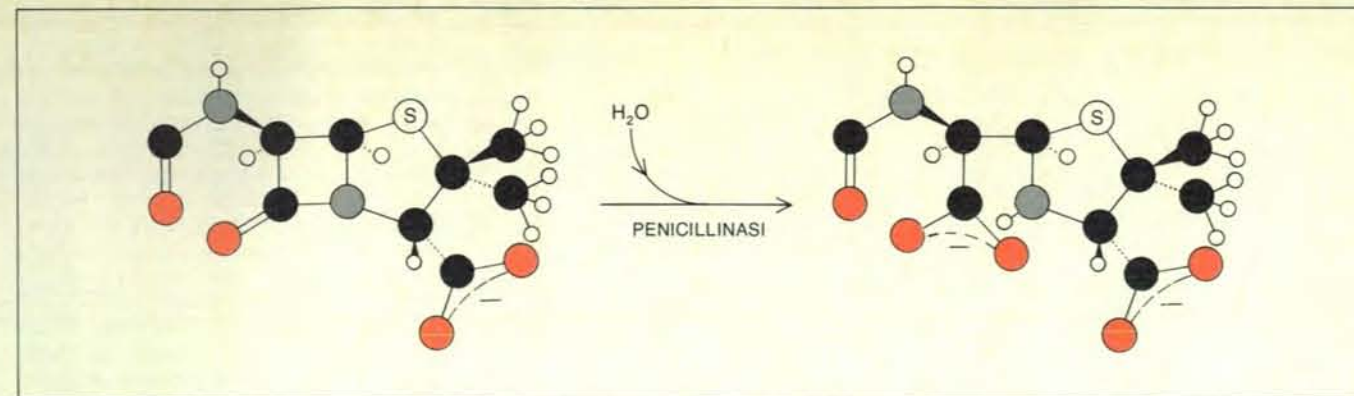
no entrambe le attività. Si ritiene che l'inibizione della formazione dei legami trasversali sia il primo della catena di eventi per cui gli antibiotici beta-lattamici portano alla morte e spesso alla lisi delle cellule batteriche. Poiché la sintesi del peptidoglicano non ha luogo nelle cellule dell'uomo e degli altri mammiferi, non è affatto sorprendente che le sostanze selezionate nel corso dell'evoluzione per inibire questo processo siano relativamente innocue per il corpo umano.

Gli eventi direttamente responsabili dell'uccisione e della lisi, tuttavia, non sono ancora compresi. Le cellule batteriche contengono sia enzimi autolitici che degradano il peptidoglicano, sia enzimi che ne permettono la sintesi. Alexander Tomasz della Rockefeller University ha scoperto che i pneumococchi contenenti una mutazione che li rende difettosi nel loro sistema autolitico non lisano in presenza della penicillina, sebbene la loro crescita venga inibita. Organismi di questo tipo vengono descritti come tolleranti. Si pone quindi la domanda se il sistema autolitico dei ceppi normali venga messo in funzione dalle modificazioni primarie causate dalla penicillina e sia quindi responsabile dell'azione litica dell'antibiotico. Il meccanismo proposto, però, è ancora incerto. Ciononostante è divenuto evidente che il meccanismo di azione degli antibiotici beta-lattamici è di considerevole complessità.

Studi recenti hanno rivelato che nella membrana citoplasmatica delle cellule batteriche esiste una molteplicità di proteine a cui gli antibiotici beta-lattamici si legano specificamente; almeno alcune di queste proteine sembrano essere enzimi con attività transpeptidasica o D-alanil carbossipeptidasica, o entrambe. Queste proteine che legano la penicillina (*penicillin-binding proteins*, PBP) possono venire identificate solubilizzando le proteine della membrana con un detergente dopo averle lasciate reagire con penicillina marcata con l'isotopo radioattivo carbonio 14; le proteine vengono successivamente separate per mezzo di una elettroforesi su gel di agar e riconosciute a causa della loro radioattività. I diversi antibiotici beta-lattamici hanno affinità differenti per una o più PBP e inducono diversi cambiamenti morfologici nei batteri. Quando agiscono sul batterio a forma di bastoncino *E. coli* a concentrazioni sufficientemente alte, la maggior parte di essi provoca la formazione di protoplasti e lisi cellulare, mentre a concentrazioni più basse alcuni di essi inducono la formazione di cellule sferiche osmoticamente stabili e altri ancora provocano la formazione di filamenti di cellule in cui la divisione cellulare è inibita.

La crescita di mutanti termosensibili di *E. coli* in cui l'una o l'altra delle PBP è stabile a 30 gradi centigradi, ma instabile a 42 gradi, è talvolta accompagnata da alterazioni morfologiche alla temperatura più alta che simulano le alterazioni causate dagli antibiotici beta-lattamici alla temperatura più bassa. Uno studio di questi mutanti compiuto da Brian G. Spratt dell'Università del Sussex ha dimostrato che la PBP designata come 1B è implicata nell'allungamento cellulare; la PBP 2 nella determinazione della forma cellulare e la PBP 3 nella divisione cellulare. L'inibizione di una qualunque di queste PBP sembra essere letale, mentre l'inibizione di altre non lo è, anche se queste ultime sono presenti in quantità molto maggiori.

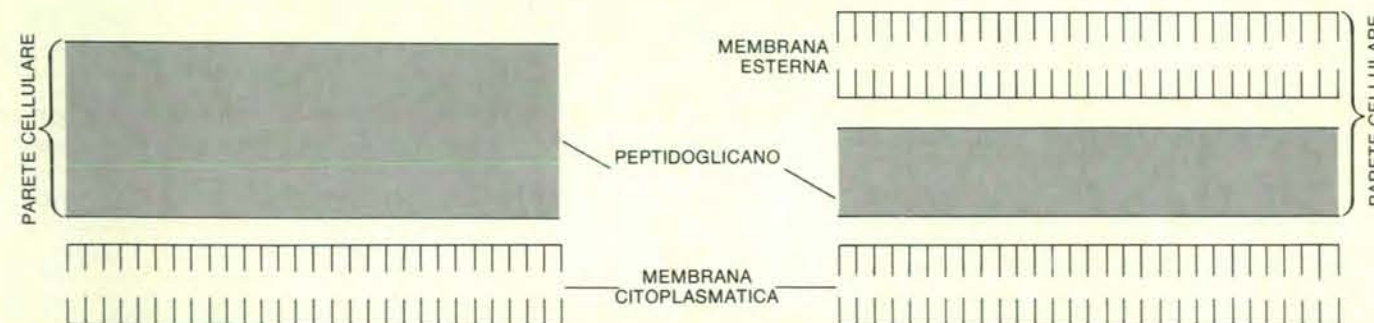
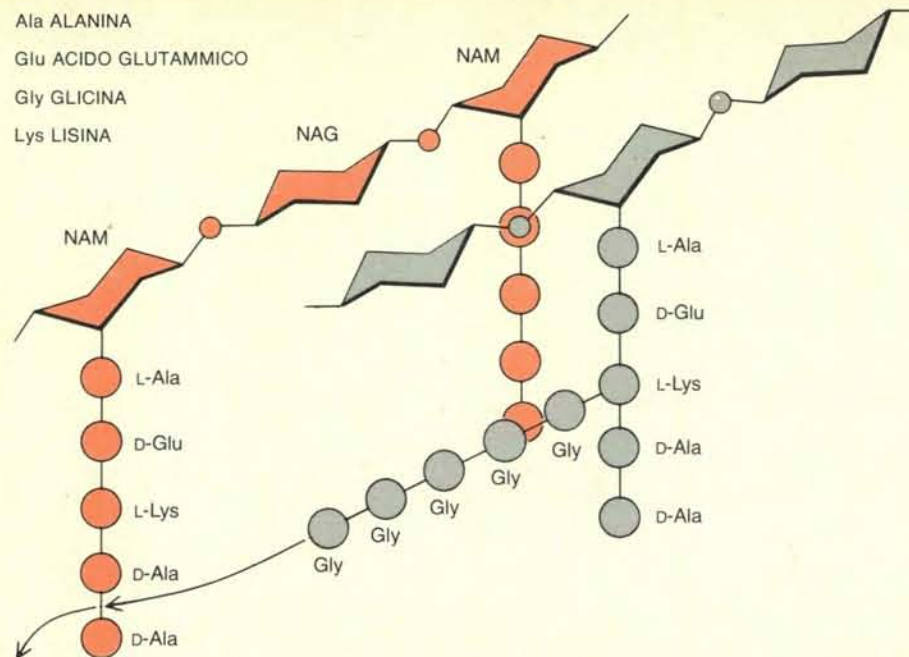
È evidente che una diminuzione della affinità dei diversi antibiotici beta-lattamici per le PBP essenziali, o una diminuzione della loro capacità di raggiungere il sito di questi enzimi, provocherà una corrispondente variazione delle loro attività antibatteriche. È stato isolato un mutante termosensibile di *E. coli* che è maggiormente resistente a due antibiotici beta-lattamici (mecillina e tienamicina) e possiede una PBP 2 con ridotta affinità per queste sostanze. È tuttavia incerto il significato clinico delle resistenze acquisite in questo modo per mutazione. La mecillina e la tienamicina esercitano i loro effetti letali su *E. coli* legandosi quasi esclusivamente alla PBP 2. Per generare un aumento di resistenza agli altri antibiotici beta-lattamici sarebbero necessarie altre mutazioni dei geni codificanti per numerose PBP. Inoltre, le modificazioni delle PBP dovrebbero essere sufficientemente specifiche da inibire il legame degli antibiotici beta-lattamici, ma non il legame dei substrati naturali richiesti per la sintesi del peptidoglicano essenziale per la sopravvivenza della cellula.

I batteri Gram-negativi tendono a essere intrinsecamente più resistenti a molti antibiotici beta-lattamici degli organismi Gram-positivi. I risultati delle ricerche sulle proprietà delle membrane esterne dei batteri Gram-negativi, intraprese da Hiroshi Nikaido dell'Università della California a Berkeley e altri, sembrano convalidare l'ipotesi che questa membrana

I batteri resistenti hanno rappresentato un problema nell'uso clinico della penicillina. Nel 1940 Ernst B. Chain e l'autore scoprirono all'Università di Oxford un enzima del batterio *Escherichia coli* che distruggeva la penicillina: lo chiamarono penicillinasi. Esso e altre penicillinasi rompono il legame ammidico dell'anello beta-lattamico della penicillina e formano l'acido penicilloico, che non possiede attività antibiotica.

contenga delle proteine (chiamate porine) che formano dei canali attraverso cui i piccoli antibiotici idrofili beta-lattamici possono, in ogni caso, avere accesso alle PBP presenti nella membrana citoplasmatica sottostante. Qualunque sia il meccanismo di penetrazione, le concentrazioni di certi antibiotici beta-lattamici richieste per inibire la crescita dei batteri Gram-negativi sono spesso più alte di quelle richieste per inattivare le PBP essenziali negli estratti cellulari, anche quando non ci siano prove di una significativa distruzione enzimatica dell'antibiotico da parte di una beta-lattamasi batterica. Al contrario, non è necessaria una concentrazione più elevata dell'antibiotico per inibire la crescita di un «mutante di permeabilità» di *E. coli*, in cui l'accesso alla membrana citoplasmatica è libero.

Sembra che la barriera rappresentata dalla membrana esterna della cellula sia almeno parzialmente responsabile della relativa insensibilità a molti antibiotici beta-lattamici di *Pseudomonas aeruginosa*, un organismo che causa gravi infezioni nei tessuti danneggiati, e della insensibilità di molti altri batteri Gram-negativi alle penicilline del tipo della meticillina. La capacità sia dei batteri Gram-negativi sia di quelli Gram-positivi di produrre beta-lattamasi che possono rompere l'anello beta-lattamico, tuttavia, è anch'essa un fattore di primaria importanza nella genesi delle resistenze e di grande significato clinico. Ciò è diventato evidente quando sono comparsi ceppi di stafilococchi produttori di penicillinasi che hanno soppiantato i ceppi sensibili. Ulteriori studi hanno dimostrato che l'enzima che inattiva la penicillina può venire in parte escreto dalle cellule e in parte rimanere attaccato a esse; inoltre la sua produzione è indotta dagli antibiotici beta-lattamici. Una conseguenza di questa situazione è che le popolazioni dense di stafilococchi, che producono una grande quantità di beta-lattamasi, distruggono rapidamente molte penicilline, sebbene le cellule singole, che ne producono solo una piccola quantità, rimangano sensibili.

Anche molti batteri Gram-negativi producono una beta-lattamasi, e in questi organismi l'enzima è spesso costitutivo, cioè viene sintetizzato in grande quantità senza che sia necessaria la presenza di un induttore. Inoltre, esso è normalmente legato alle cellule, sebbene si trovi spesso nello spazio periplasmatico tra la membrana citoplasmatica e la parete cellulare esterna e venga facilmente rilasciato quando la parete viene danneggiata. Quando le beta-lattamasi, ottenute da batteri diversi, furono «saggiate» con una serie di penicilline e cefalosporine divenne evidente che esistevano molti enzimi di questa famiglia che differivano tra loro per la capacità di idrolizzare i diversi antibiotici beta-lattamici.

La molteplicità delle beta-lattamasi rappresenta un problema serio per il medico, così come la scoperta che i geni che specificano per la produzione delle beta-lattamasi possono venire trasferiti da una specie batterica all'altra. Questi geni non sono essenziali per la vita dei batteri e si trovano spesso sulle unità genetiche autoreplicanti note come plasmidi. Richard P. Novick del Public Health Research Institute of the City of New York e Mark H. Richmond dell'Università di Bristol hanno dimostrato che la capacità di produrre penicillinasi negli stafilococchi è mediata da plasmidi che possono venire trasferiti da uno stafilococco all'altro, dopo essere stati incorporati nella testa di un virus che infetta la cellula. Il trasferimento per mezzo di questo meccanismo, noto come trasduzione, è probabilmente di poca importanza clinica. Nei batteri Gram-negativi, tuttavia, i plasmidi possono essere trasferiti da una cellula all'altra della stessa specie o di una specie differente per mezzo di un processo noto come coniugazione, in cui una

La struttura della parete cellulare dei batteri Gram-positivi (*a sinistra*) e Gram-negativi (*a destra*) contribuisce a spiegare il fatto che la penicillina *G* è attiva contro i batteri Gram-positivi, ma non contro quelli Gram-negativi. In entrambi i tipi la membrana citoplasmatica è ricoperta da uno strato di peptidoglicano, ma i batteri Gram-negativi possiedono una membrana esterna e possono inoltre immagazzinare gli enzimi beta-lattamasi nello spazio periplasmatico tra la membrana citoplasmatica e lo strato di peptidoglicano. La membrana citoplasmatica contiene proteine che legano la penicillina; queste, essendo coinvolte nella sintesi della parete cellulare, sono vulnerabili all'azione della penicillina.

La struttura del peptidoglicano della parete cellulare batterica consiste di unità alternate di due amminozuccheri, la N-acetilglucosammina (NAG) e l'acido N-acetilmuramico (NAM). Le unità di NAM sono legate alle catene peptidiche, molte delle quali presentano anche legami incrociati tra di loro. Nel batterio *Staphylococcus aureus* i legami incrociati vengono realizzati inserendo una glicina terminale nel legame tra le due unità di D-alanina di un'altra catena. Gli enzimi transpeptidasici che realizzano questa inserzione (quando i batteri stanno moltiplicandosi e quindi costruendo nuova parete) possono venire attaccati dalle penicilline e dalle cefalosporine.

cellula donatrice «maschio» aderisce per breve tempo a una cellula ricevente «femmina». Inoltre i geni che codificano per la beta-lattamasi possono essere presenti in certi frammenti specializzati di DNA, noti come trasposoni, che possono venire trasferiti da un plasmide all'altro.

La capacità di una cellula batterica o di un plasmide di infettare un'altra cellula del loro tipo con geni che specificano per una beta-lattamasi costituisce un meccanismo potenzialmente importante di diffusione della resistenza. La resistenza può essere trasferita da un organismo patogeno all'altro, ma anche da un organismo non patogeno a uno patogeno. In alcuni casi, tuttavia, la resistenza può richiedere un certo tempo per manifestarsi. Virtualmente tutti i gonococchi sono rimasti sensibili alla benzilpenicillina per molti anni, finché è stato identificato, in alcuni pazienti affetti da gonorrea, un ceppo resistente portatore di una beta-lattamasi codificata da un plasmide e simile a una precedentemente nota in altri batteri Gram-negativi.

In queste condizioni, qual è stato il successo della ricerca di nuovi antibiotici beta-lattamici in grado di risolvere i problemi clinici? Un risultato significativo è derivato dalla scoperta avvenuta a Oxford nel 1953 che un fungo del genere *Cephalosporium*, isolato nel 1945 da G. Brotzu in Sardegna, produceva una nuova penicillina con una catena laterale derivante dall'acido D-alfa-amminoadipico. G. G. F. Newton e io scoprimmo che preparati grezzi di questa penicillina contenevano piccole quantità di un secondo antibiotico con la stessa catena laterale, ma in possesso di una caratteristica strutturale fino ad allora mai osservata: l'anello beta-lattamico era fuso a un anello a sei membri invece che a uno a cinque membri. Il nuovo antibiotico, chiamato cefalosporina *C*, si dimostrò di potenziale interesse medico in quanto possedeva la caratteristica bassa tossicità delle penicilline verso gli animali, ma era resistente all'idrolisi da parte di certe beta-lattamasi.

Quando *Penicillium* e *Cephalosporium* vengono fatti crescere in laboratorio o in impianti industriali, richiedono condizioni di fermentazione marcatamente diverse. Entrambi i funghi, tuttavia, seguono vie biosintetiche simili nella produzione degli antibiotici beta-lattamici. Gli studi eseguiti con estratti cellulari hanno dimostrato che una catena di tre amminoacidi viene trasformata in un anello per dare origine all'isopenicillina *N*, una penicillina con una catena laterale L-alfa-amminoadipile. Nel *Penicillium chrysogenum* una aciltransferasi catalizza poi la sostituzione di questa catena laterale con un'altra derivata da un precursore utilizzabile, come l'acido fenilacetico. Nelle specie di *Cephalosporium* la transferasi sembra essere assente, dato che non si forma alcun composto con una nuova catena laterale. Sono invece necessarie due reazioni sequenziali. Il centro asimmetrico della catena laterale L-alfa-amminoadipile viene convertito nella struttura tridimensionale speculare. Parte della penicillina *N* risultante, con una catena laterale di acido D-alfa-amminoadipico, viene quindi escreta, ma recenti esperimenti di Arnold L. Demain e dei suoi colleghi al MIT indicano che una parte della penicillina *N* viene convertita in deacetossicefalosporina *C*, che è il precursore della cefalosporina *C* e di altre cefalosporine.

Per mezzo delle sole fermentazioni è stato possibile ottenere determinati tipi di penicilline e solo le cefalosporine con la catena laterale D-alfa-amminoadipile. Tuttavia con metodi semisintetici è stato possibile produrre e studiare migliaia di membri della famiglia degli antibiotici beta-lattamici. La semisintesi è divenuta realizzabile con la scoperta dei metodi per la produzione del nucleo della molecola di penicillina, l'acido 6-amminopenicillanico (6-APA), e della molecola di cefalosporina, l'acido 7-alfa-amminocefalosporanico (7-ACA), in cui è assente una catena laterale acilica. In seguito all'annuncio della produzione di 6-APA in Giappone dal *Penicillium chrysogenum*, questa sostanza fu sintetizzata da John Sheehan e scoperta indipendentemente nei Beecham Research Laboratories in Gran Bretagna, dove fu isolata da fermentazioni in cui non era stato aggiunto alcun precursore per la catena laterale. In seguito fu prodotta in grande

Le proteine che legano la penicillina (PBP) delle membrane citoplasmatiche del batterio *E. coli* furono identificate con le tecniche elettroforetiche applicate da Brian G. Spratt dell'Università del Sussex. La penicillina marcata con l'isotopo radioattivo carbonio 14 fu fatta reagire con le membrane. Le proteine furono poi frazionate mediante elettroforesi su gel che fa migrare le proteine a velocità differenti a seconda del peso molecolare. Le fotografie mostrano il gel colorato per evidenziare le proteine *(a)*, le proteine marcate identificate con l'autoradiografia *(b)* e la separazione delle proteine che legano la penicillina 1A e 1B con metodi elettroforetici a risoluzione migliore *(c)*.

La biosintesi delle penicilline e delle cefalosporine prodotte dai funghi *Penicillium* e *Cephalosporium* inizia *(1)* con la formazione di un tripeptide; da sinistra a destra le tre unità, separate da linee tratteggiate, sono l'acido L-alfa-amminoadipico, la L-cisteina e la L-valina. Il tripeptide viene convertito in isopenicillina *N (2)*. Nelle penicilline un enzima transferasi catalizza la sostituzione della catena laterale L-alfa-amminoadipile dell'isopenicillina *N (3)*, mentre nelle cefalosporine la transferasi è apparentemente assente, dato che non si forma alcun composto con una nuova catena laterale. Avvengono, invece, due reazioni sequenziali. Per prima cosa il centro asimmetrico della catena laterale L-alfa-amminoadipile viene invertito nella sua immagine tridimensionale speculare per formare la penicillina *N*. Successivamente la penicillina *N* viene convertita in deacetossicefalosporina facendo espandere l'anello: si rompe il legame tra l'atomo di zolfo e il secondo carbonio e si stabilisce un nuovo legame tra il carbonio del gruppo beta-metilico e lo zolfo. I meccanismi delle reazioni enzimatiche non sono ancora chiariti.

quantità mediante rimozione della catena laterale della benzilpenicillina con acilasi purificate da batteri.

Non fu invece possibile trovare alcun enzima capace di rimuovere la catena laterale D-alfa-amminoadipile dalla cefalosporina *C*. Malgrado ciò, il 7-ACA fu prodotto in piccole quantità a Oxford idrolizzando in condizioni debolmente acide la cefalosporina *C* e in seguito per mezzo di una procedura chimica molto più efficiente ideata dagli Eli Lilly Research Laboratories. Questi risultati spianarono la strada alla preparazione di penicilline e cefalosporine semisintetiche attraverso l'aggiunta di catene laterali differenti al 6-APA e al 7-ACA. Inoltre, la molecola di cefalosporina poté essere ulteriormente modificata senza alterare il suo essenziale sistema ad anello per mezzo di procedure chimiche che scambiano il gruppo acetossi (-$OCOCH_3$) legato all'atomo di carbonio designato 3' e sporgente dall'anello a sei membri con altri gruppi o con un atomo di idrogeno.

Due altre scoperte ebbero rilevanza nel quadro delle ricerche sulle cefalosporine. La prima fu l'invenzione nei laboratori della Lilly di una ingegnosa via di sintesi chimica a partire dal sistema ad anelli della penicillina da quello della cefalosporina, che permise la trasformazione di alcune penicilline nelle cefalosporine corrispondenti. La seconda fu la scoperta nei laboratori della Lilly e della Merck Sharp & Dohme che le cefalosporine prodotte dagli *Streptomyces* contenevano un gruppo metossilico (-OCH_3) al posto di un idrogeno attaccato al carbonio che si trova in diagonale, nell'anello beta-lattamico, rispetto all'atomo di azoto. In seguito furono sviluppati dei metodi per introdurre il gruppo metossilico nelle altre molecole di cefalosporina.

Molte migliaia di penicilline e cefalosporine sono state sintetizzate finora con l'uno o l'altro di questi metodi, e alcune di esse hanno trovato un posto importante nella pratica medica. La prima delle nuove penicilline clinicamente utili, chiamata meticillina, fu sintetizzata nei Beecham Laboratories e si dimostrò efficace contro gli stafilococchi benzilpenicillina-resistenti, principalmente a causa del fatto che la sua voluminosa catena laterale la proteggeva dall'azione delle penicillinasi stafilococciche. Successivamente vennero l'ampicillina, più attiva contro un certo numero di batteri Gram-negativi, e la carbenicillina, efficace contro il batterio *Pseudomonas aeruginosa*.

Le prime due cefalosporine semisintetiche, la cefalotina della Lilly e la cefaloridina dei Glaxo Laboratories in Gran Bretagna, erano attive anche contro gli stafilococchi penicillina-resistenti; in questo caso era la natura del sistema ad anelli, piuttosto che la catena laterale, a proteggere gli antibiotici da una rapida idrolisi da parte dell'enzima stafilococcico. Le due nuove cefalosporine si dimostravano efficaci anche contro un certo numero di batteri Gram-negativi e inoltre non provocavano alcuna reazione nella maggior parte dei pazienti ipersensibili alla penicillina. Sebbene il sistema ad anelli della cefalosporina sia piuttosto stabile nelle condizioni acide dello stomaco, le prime cefalosporine non venivano assorbite se somministrate per via orale. Tuttavia, inaspettatamente, un'altra cefalosporina semisintetica, la cefalexina, si dimostrò efficace anche per via orale.

Malgrado le favorevoli proprietà di queste prime cefalosporine e penicilline semisintetiche, alcuni batteri patogeni Gram-negativi sono insensibili alla loro azione. Altri cambiamenti dei gruppi legati all'anello beta-lattamico hanno ora prodotto una seconda e una terza generazione di cefalosporine. I primi tre prodotti furono chiamati cefamandolo, cefuroxima e cefoxitina e furono presto seguiti dalla cefotaxima e dal cefoperazone. Inoltre per semisintesi dalla penicillina è stata ottenuta una 1-oxacefalosporina altamente attiva, in cui un atomo di ossigeno sostituisce un atomo di zolfo nell'anello a sei membri. Alcuni di questi composti rappresentano progressi notevoli nella chemioterapia antibatterica perché possiedono uno spettro di attività più ampio degli antibiotici beta-lattamici precedenti e sono resistenti all'idrolisi da parte di una grande varietà di beta-lattamasi.

Un nuovo capitolo nella storia degli antibiotici beta-lattamici è stato aperto nel 1976 quando alcuni laboratori farmaceutici hanno scoperto che certi *Streptomyces* producono sostanze che contengono l'anello beta-lattamico, ma sono privi del sistema ad anelli della penicillina e di quello della cefalosporina. Alcune di queste nuove sostanze sono state scoperte durante *screening* di sostanze antibatteriche; altre sono state trovate durante screening di sostanze capaci di inibire le beta-lattamasi. Una di esse, la tienamicina, ha un ampio spettro antibatterico e una alta resistenza alle beta-lattamasi e potrà verosimilmente trovare un'utilizzazione clinica. Un'altra, l'acido clavulanico, è un potente inattivatore delle beta-lattamasi. In questo stesso periodo Robert Woodward ha descritto la sintesi e la semisintesi di composti beta-lattamici reattivi noti come penemi, in cui il doppio legame dell'anello contenente zolfo delle cefalosporine è stato introdotto nell'anello corrispondente delle penicilline. Ancora più recentemente altri nuovi antibiotici beta-lattamici sono stati isolati dai terreni di coltura di molte specie di batteri.

Fino a che punto è stato possibile prevedere che tipo di antibiotici beta-lattamici era necessario sintetizzare per tentare di risolvere i problemi della pratica clinica? Alcuni punti fondamentali sono stati individuati rapidamente. Fu presto noto che una catena laterale capace di attirare un elettrone era in grado di rendere una penicillina sufficientemente stabile in ambiente acido da poter essere somministrata con successo per via orale.

I nuovi antibiotici beta-lattamici, scoperti dopo il 1976, comprendono la tienamicina *(a)*, l'acido clavulanico *(b)* e i penemi, uno dei quali è raffigurato in *c*. La tienamicina e l'acido clavulanico sono prodotti dai batteri del gruppo *Streptomyces*. Questi antibiotici possiedono tutti un anello beta-lattamico, ma sono privi sia del sistema ad anelli della penicillina sia di quello della cefalosporina. I penemi, sintetizzati dal defunto Robert B. Woodward della Harvard University, sono invece caratterizzati dalla presenza del doppio legame nell'anello della penicillina nella posizione corrispondente al doppio legame dell'anello contenente zolfo delle cefalosporine.

Le cefalosporine della terza generazione comprendono la cefotaxima *(in alto)* e il moxalattame *(in basso)*, una 1-oxacefalosporina altamente attiva derivata dalla penicillina per semisintesi. Una caratteristica del secondo è l'atomo di ossigeno che sostituisce l'azoto nell'anello a sei membri.

Per analogia con le differenze note tra la penicillina *N* e la benzilpenicillina sembrava verosimile che la sostituzione della catena laterale alfa-amminoadipile delle cefalosporine naturali con catene laterali meno polari (cioè catene laterali senza cariche positive o negative) avrebbe prodotto un notevole aumento dell'attività intrinseca contro alcuni batteri e il mantenimento della resistenza alla penicillinasi stafilococcica.

In effetti fu così. La scoperta che il gruppo O-acetile sul carbonio 3' delle cefalosporine, che ho già citato, poteva essere rimosso nell'organismo da un'acetilesterasi, con la conseguente produzione di composti meno attivi, suggerì che questo gruppo potesse essere sostituito con maggiore profitto da gruppi più stabili. Tuttavia, le caratteristiche strutturali richieste per l'assorbimento intestinale delle cefalosporine non avrebbero potuto essere previste, così come non avrebbero potuto essere previste le caratteristiche richieste perché un antibiotico beta-lattamico si dimostrasse allo stesso tempo un inattivatore potente di enzimi essenziali delle membrane delle cellule batteriche e un composto resistente all'idrolisi da parte di molte beta-lattamasi. Così, sebbene siano stati sintetizzati molte migliaia di composti, solo alcuni di questi si sono dimostrati utili nella pratica medica. Questi ultimi sono stati ottenuti principalmente per mezzo dello sfruttamento di osservazioni casuali e dalla scoperta di nuovi tipi di molecole beta-lattamiche in ampi e sistematici programmi di screening.

Una delle ragioni di questa situazione è la mancanza di conoscenze sui siti attivi tridimensionali dei vari enzimi con cui gli antibiotici beta-lattamici reagiscono. Significativi progressi nella nostra comprensione della natura e della funzione di questi siti stanno ora per essere raggiunti.

Le proteine che legano la penicillina e che possiedono un'attività D-carbossipeptidasica possono essere purificate per mezzo della cromatografia per affinità, in cui una penicillina o cefalosporina adatta viene legata a una resina. Le PBP si legano specificamente alla colonna cromatografica e possono venire eluite con idrossilammina, che rompe il legame covalente formato dalla reazione del PBP con l'anello beta-lattamico. Adottando questo procedimento sono state determinate le sequenze amminoacidiche di frammenti di due tra le PBP principali ed è stata individuata l'unità amminoacidica tramite cui esse si legano alla penicillina. La purificazione delle PBP minori, che presumibilmente sono delle transpeptidasi inattivabili dagli antibiotici beta-lattamici, si è dimostrata difficile; recentemente sono stati ottenuti dei progressi inserendo in *E. coli* sequenze di DNA che inducono la produzione di copie multiple del gene che specifica per una particolare PBP. Ciò è stato realizzato inserendo i geni in un plasmide presente in numerose copie all'interno della cellula o in un virus che produce molte copie del gene quando infetta la cellula batterica e si moltiplica.

La purificazione di molte beta-lattamasi è già stata realizzata. Richard P. A. Ambler dell'Università di Edimburgo ha determinato le sequenze amminoacidiche degli enzimi prodotti da *Staphylococcus aureus*, *Bacillus licheniformis* e *B. cereus*, mentre J. G. Sutcliffe di Harvard ha dedotto la sequenza di una beta-lattamasi di un plasmide di *E. coli* dalla sequenza nucleotidica del gene corrispondente. A parte un'unica eccezione (un enzima richiedente zinco prodotto da *B. cereus*), queste sequenze mostrano un'omologia considerevole, cioè sono molto più simili di quanto ci si potrebbe aspettare per effetto del caso. È stata anche riportata l'esistenza di significative omologie tra le beta-lattamasi e due PBP con attività D-carbossipeptidasica. È perciò possibile che un gruppo di beta-lattamasi e un gruppo di enzimi delle membrane batteriche sensibili alla penicillina abbiano avuto un'origine evolutiva comune.

Queste due famiglie enzimatiche sono simili anche nel loro meccanismo di reazione. Per mezzo dell'uso di composti simili alla benzilpenicillina che rimangono legati ai siti attivi è stato dimostrato che il gruppo idrossile di un particolare amminoacido, la serina, si attacca all'a-

nello beta-lattamico. Ciò provoca l'apertura dell'anello e la formazione di un acil-enzima che è un estere della serina. Nel caso di una beta-lattamasi, l'acil-enzima prodotto da molti antibiotici beta-lattamici viene rapidamente idrolizzato, rigenerando l'enzima e liberando un penicilloato o un composto analogo senza alcuna attività antibatterica. Nel caso di una carbossipeptidasi della membrana batterica la scissione dell'acil-enzima prodotto dagli antibiotici beta-lattamici è lenta, così che l'acil-enzima inattivo si accumula.

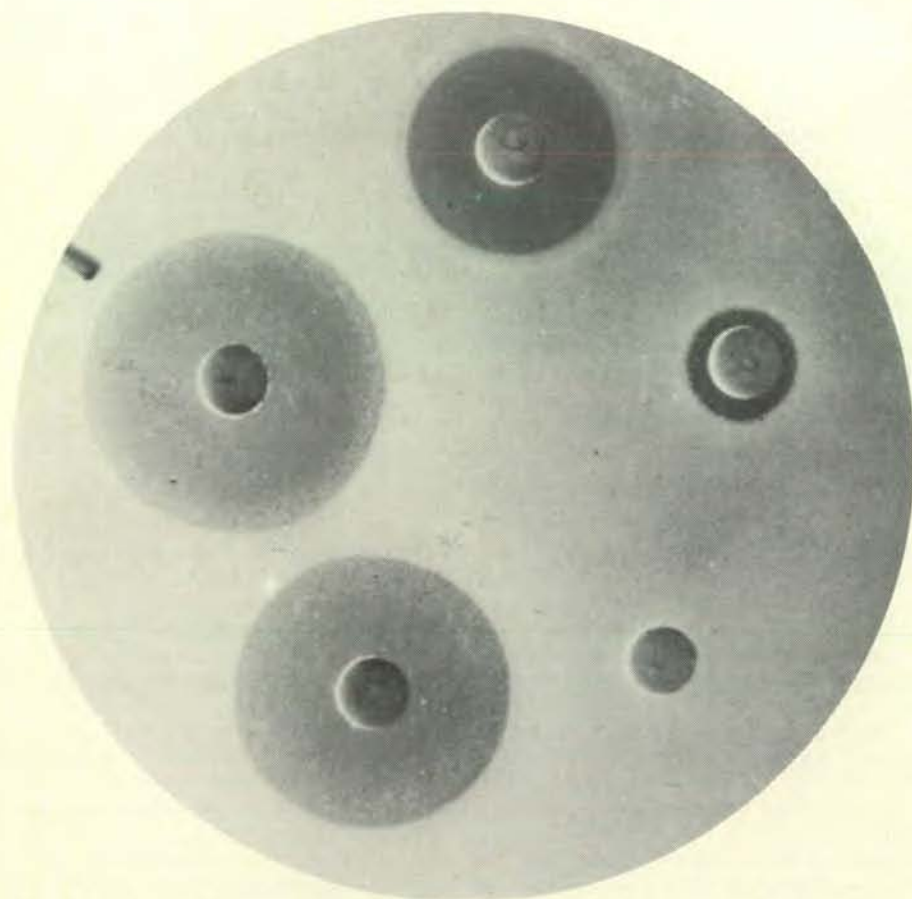
In una D-carbossipeptidasi è stato dimostrato che l'unità di serina che reagisce con la penicillina è lo stesso amminoacido che reagisce con il substrato dell'enzima contenente D-alanil-D-alanina. Questo è in accordo con un'ipotesi avanzata nel 1965 da Jack Strominger e Donald J. Tipper dell'Università del Wisconsin a Madison. Per spiegare perché la penicillina inibisce la formazione di legami incrociati durante la sintesi della parete cellulare, essi postularono che la localizzazione spaziale di una sequenza di atomi della struttura della penicillina simulasse la localizzazione spaziale dei corrispondenti atomi delle estremità terminali D-alanil-D-alanina delle catene laterali nascenti del peptidoglicano. Quanto questa ipotesi possa essere estesa, per spiegare le potenzialità relative dei diversi antibiotici beta-lattamici nell'inibizione della funzione delle PBP di membrana essenziali, non è dato sapere.

Tra i nuovi composti beta-lattamici molto più resistenti della benzilpenicillina all'idrolisi da parte di enzimi della famiglia delle beta-lattamasi vi sono alcuni la cui interazione con una beta-lattamasi provoca una inattivazione temporanea o permanente dell'enzima. Questo fenomeno viene studiato attualmente da Jeremy Knowles a Harvard, Stephen G. Waley a Oxford e molti altri. Se la velocità di idrolisi dell'acil-enzima intermedio è più lenta di quella della sua formazione, l'intermedio si accumula e il sito attivo non è più disponibile per l'idrolisi di un substrato facilmente idrolizzabile come la benzilpenicillina. Dopo che il substrato lentamente idrolizzabile è stato utilizzato completamente e non è più disponibile per formare nuovo acil-enzima, tuttavia, l'enzima attivo può venire rigenerato. In alcuni casi, d'altra parte, il componente acilico delle molecole di acil-enzima subisce una modificazione strutturale che ne impedisce completamente la rimozione. Quando questo accade, si dice che l'enzima si è «suicidato».

Il concetto che un antibiotico beta-lattamico dovrebbe proteggere un secondo antibiotico dalla distruzione da parte delle beta-lattamasi nel corso della terapia risale al 1956, quando si scoprì che l'idrolisi della benzilpenicillina da parte di una beta-lattamasi era inibita competitivamente dalla cefalosporina *C*. La somministrazione di miscele di due composti sinergici per questo motivo non ha ricevuto molto favore. Ciononostante, in seguito alla scoperta dell'acido clavulanico e di altre sostanze che non solo inibiscono le beta-lattamasi competitivamente, ma anche le inattivano irreversibilmente, le potenzialità di questo sinergismo saranno ulteriormente studiate.

Gli incredibili sviluppi nel campo degli antibiotici beta-lattamici in seguito alla scoperta dell'attività terapeutica della penicillina e l'attuale intensa ricerca in questo campo sono stati stimolati in misura non certo trascurabile dall'enorme utilità medica di questi antibiotici non tossici. Sono stati realizzati progressi sostanziali verso la soluzione dei problemi rappresentati dai batteri resistenti alla penicillina, sebbene i meccanismi in rapida evoluzione della resistenza e la grande mutabilità dei microrganismi facciano ritenere poco saggio considerare queste soluzioni come definitive.

Trent'anni fa l'attività della molecola di penicillina sembrava essere dovuta alle proprietà del suo singolare sistema ad anelli fusi. Oggi l'unica caratteristica comune nella struttura degli antibiotici beta-lattamici noti è l'anello beta-lattamico stesso. L'atomo di azoto dell'anello è comunemente legato a un atomo di carbonio che porta un gruppo carbossile (-COOH), un gruppo ionizzabile correlato in forma simile o un lattone (un estere ciclico di un acido idrossilico), ma nelle nuove beta-lattamine ottenute dai batteri, questo atomo di azoto fa parte di un gruppo sulfammato ($-NSO_3^-$). Perché la molecola possieda un'alta attività intrinseca contro i batteri sembra necessario che la beta-lattamina sia altamente reattiva. Questo si ottiene nelle penicilline e nelle cefalosporine a causa della fusione dell'anello beta-lattamico con un secondo anello per generare un sistema in cui la configurazione normale è deformata e sottoposta a una tensione.

Se la beta-lattamina è troppo reattiva, tuttavia, la molecola non sarà abbastanza stabile da poter essere usata per scopi clinici, e il suo effetto biologico non sarà più altamente selettivo. Inoltre, se i gruppi legati all'anello non occupano le posizioni idonee nello spazio tridimensionale, la molecola non si adatta ai siti attivi delle proteine della membrana cellulare che legano la penicillina e non forma un'acil-enzima relativamente stabile. Nei prossimi dieci anni potrebbe essere possibile specificare le caratteristiche di una beta-lattamina potenzialmente utile con precisione molto maggiore.

L'effetto dei penemi è indicato in questa fotografia che mostra l'attività inibitoria di numerosi antibiotici contro i batteri resistenti alla penicillina della specie *Staphylococcus aureus* fatti crescere su agar. Ognuno dei cinque cerchi scuri rappresenta un'area in cui i batteri sono stati uccisi da un particolare antibiotico; più il cerchio è grande, più è efficace l'antibiotico. In senso orario a partire dall'alto gli antibiotici sono la cefaleina; la penicillina *V*; un peneme inattivo, in cui l'atomo di zolfo è disposto al di sotto del piano dell'atomo di carbonio beta-lattamico invece che sopra come nelle penicilline e nelle cefalosporine, producendo una forma tridimensionale non adatta; una miscela di penemi che presentano una commistione di alcune configurazioni naturali e altre no; e, infine, un peneme attivo con una configurazione tridimensionale naturale.

Reattori nucleari raffreddati a gas

Gli Stati Uniti hanno in esercizio una sola di queste unità che presentano l'interessante caratteristica di non consentire un incidente di perdita di refrigerante come quello verificatosi a Three Mile Island

di Harold M. Agnew

Nel marzo 1979 l'industria elettronucleare ha subito uno shock dal quale non si è ancora riavuta, in seguito all'incidente che ha messo fuori uso uno dei reattori nucleari di Three Mile Island. Per ironia della sorte, quell'evento, che non causò danni fisici percettibili ad alcuno, paralizzò il prospettato sviluppo dell'energia nucleare proprio nel momento in cui gli Stati Uniti acquisivano una generale consapevolezza dell'indispensabilità di nuove fonti nazionali di energia.

L'esperienza di Three Mile Island anche se ha dimostrato, con il compiacimento dei tecnici, che gli attuali reattori nucleari raffreddati ad acqua non costituiscono una minaccia per la salute e la sicurezza della popolazione, ha messo tuttavia in evidenza che incidenti e malfunzionamenti del genere sono in grado di porre a repentaglio l'esercizio dell'impianto, sottoponendo conseguentemente a un grave rischio il relativo cospicuo investimento finanziario. Nel caso limite, un incidente del tipo di Three Mile Island potrebbe minacciare la stessa sopravvivenza economica della società produttrice di elettricità.

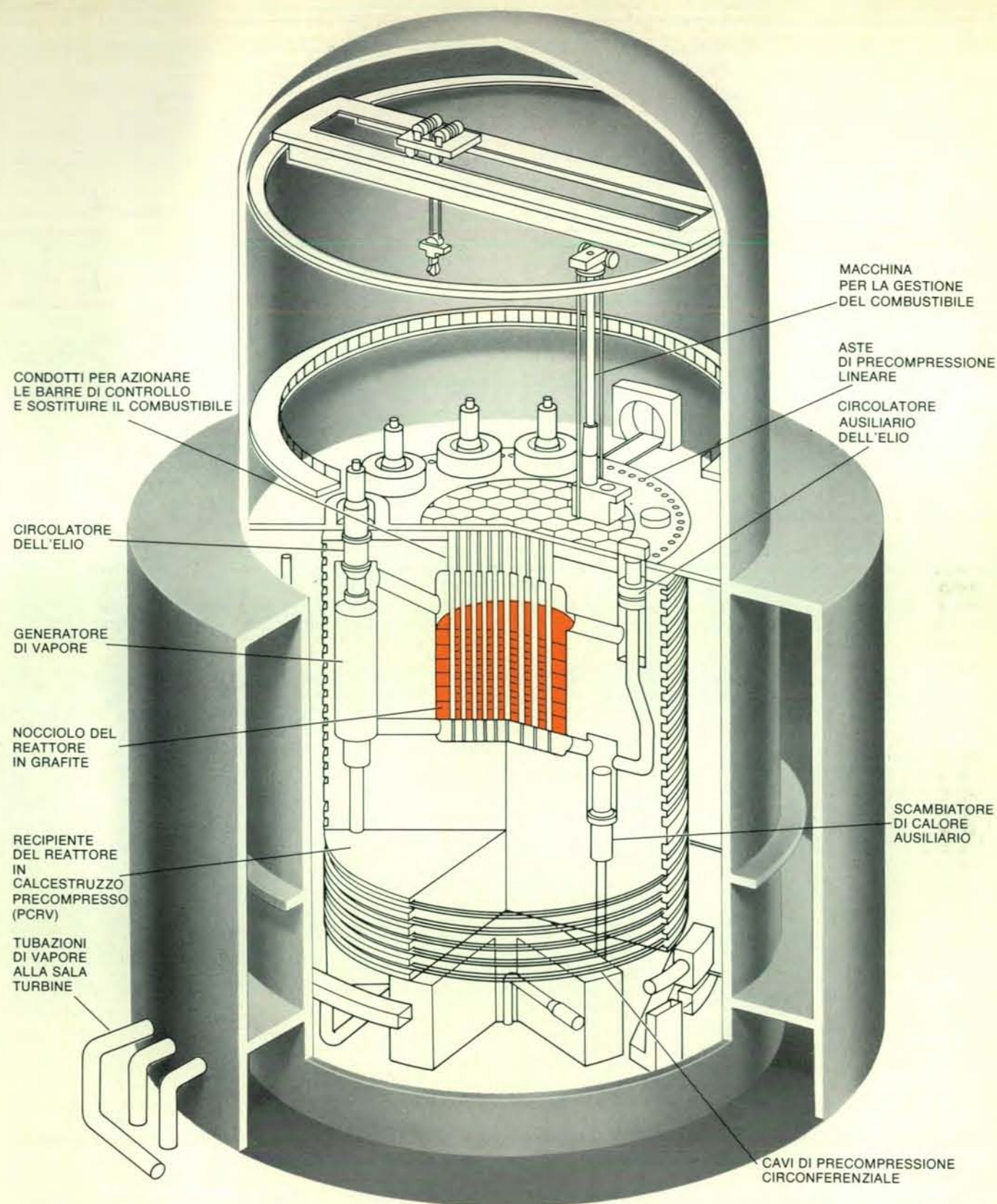
La lezione principale di Three Mile Island è forse questa: gli attuali impianti elettronucleari di potenza sono soggetti a taluni eventi molto rari che possono dar luogo a situazioni in cui il tempo a disposizione per una corretta risposta può essere inferiore al minuto. Per simili eventi a bassa probabilità, se gli interventi appropriati non vengono effettuati immediatamente, possono verificarsi conseguenze gravissime, pur senza rischi per la sicurezza pubblica. È quindi giusto chiedersi se ci si deve ritenere soddisfatti di reattori nucleari il cui progetto comporta che gli operatori debbano prendere le corrette decisioni nello spazio di un minuto per prevenire danni all'impianto. E la risposta è negativa.

Perché l'industria nucleare statunitense ha sviluppato impianti con i limiti cui si è accennato? Il predominio negli Stati Uniti dei reattori ad acqua naturale si spiega molto semplicemente. I reattori del tipo ad acqua in pressione sono un adattamento diretto del reattore compatto progettato per la propulsione del primo sommergibile nucleare, il *Nautilus*, varato nel 1954. Una versione del reattore per sommergibili adattata alla generazione elettrica di potenza venne realizzata dalla Westinghouse Eletric Company ed entrò in servizio tre anni dopo a Shippingport, Pennsylvania. Ben presto la General Electric Company introdusse sul mercato un proprio impianto ad acqua bollente, in cui il calore prodotto dalla fissione nucleare viene rimosso dal nocciolo da vapore anziché da acqua in pressione.

In entrambi i tipi di reattore è indispensabile che il nocciolo non venga a trovarsi scoperto, neppure per tempi brevi, altrimenti la temperatura nel nocciolo aumenta rapidamente provocando la fusione dell'incamiciatura metallica delle pastiglie di combustibile, come probabilmente è avvenuto a Three Mile Island. I reattori ad acqua naturale sono equipaggiati con dispositivi di sicurezza ridondanti per potere adeguatamente far fronte a un incidente di «perdita di refrigerante». In questo tipo di incidente i sistemi di emergenza intervengono ad allagare il nocciolo con acqua erogata da fonte sicura per quantità e durata. Nel caso di Three Mile Island, quando si interruppe il normale flusso di acqua di raffreddamento, una sequenza di eventi improbabili, tra cui interventi chiaramente errati degli operatori, provocò l'interruzione della gettata d'acqua d'emergenza per un periodo di tempo eccessivo.

Tutti i 71 impianti elettronucleari di potenza in servizio commerciale negli Stati Uniti, eccetto uno, sono del tipo ad acqua naturale e attualmente producono circa l'11 per cento del totale nazionale di energia elettrica. L'eccezione è rappresentata da un reattore raffreddato a elio, con cui è equipaggiata la centrale nucleare di Fort St. Vrain, in servizio commerciale dall'estate del 1979 e di proprietà della Public Service Company of Colorado. La potenza nominale dell'impianto è di 330 megawatt (MWe), corrispondente a circa un terzo di quella degli impianti elettrogeneratori standard. Il reattore ha funzionato fino al livello del 70 per cento della potenza nominale e recentemente la Nuclear Regulatory Commission ha autorizzato le prove a piena potenza.

L'impianto dimostrativo di Fort St. Vrain è stato costruito per la Public Service Company of Colorado dalla General Atomic Company, nell'ambito del Programma dimostrativo reattori di potenza, della Atomic Energy Commission. L'iniziativa faceva seguito al successo di un prototipo da 40 MWe, l'unità n. 1 della centrale nucleare di Peach Bottom, della Philadelphia Electric Company. Nel corso dei suoi sette anni e mezzo di esercizio, dal 1967 al 1974, il reattore di Peach Bottom fece registrare un fattore di disponibilità in servizio dell'86 per cento (esclusi i periodi di arresto programmato per attività di ricerca e sviluppo relativa a questo tipo di reattore). Il corrispondente valore per tutti gli altri reattori statunitensi è intorno al 66 per cento.

Due sono le caratteristiche di sicurezza che differenziano il reattore raffreddato a elio dai reattori raffreddati ad acqua. Innanzitutto il nocciolo viene raffreddato da gas che circola completamente confinato all'interno di un massiccio recipiente a pressione e, quindi, il reattore non può perdere il proprio fluido refrigerante primario in seguito alla rottura di una tubazione all'esterno del recipiente. In secondo luogo, se la circolazione del gas viene a interrompersi per un guasto all'intero sistema principale di circolazione dell'elio, la temperatura all'interno del nocciolo del reattore aumenta solo lentamente, in quanto gli elementi di combustibile sono immersi in una massiccia matrice di grafite che funge da moderatore per il rallenta-

Il reattore a gas ad alta temperatura (HTGR, High Temperature Gas-cooled Reactor) è un concetto di impianto di seconda generazione più efficiente dei 71 impianti ad acqua naturale che attualmente forniscono circa l'11 per cento dell'energia elettrica degli Stati Uniti. In questo HTGR progettato dalla General Atomic Company il moderatore (il materiale che rallenta i neutroni nel nocciolo del reattore) è grafite e il fluido refrigerante elio. Nei reattori ad acqua naturale, questa (opportunamente demineralizzata e condizionata) funge sia da moderatore sia da refrigerante. L'impianto qui illustrato ha una potenza di 860 megawatt elettrici (MWe), un po' inferiore rispetto agli impianti di massima potenza, da 1000 o più MWe. Il rendimento termico di un impianto HTGR è del 38,5 per cento, dell'ordine di quello dei migliori impianti termoelettrici a combustibile fossile e superiore a quello degli attuali reattori ad acqua naturale (32-33 per cento). Se si verifica l'interruzione o una perdita del flusso refrigerante, un reattore HTGR è molto meno esposto al rischio di danneggiamento rispetto a un reattore ad acqua naturale, grazie al nocciolo contenente circa 1500 tonnellate di grafite, che ha una grande capacità di assorbimento del calore. L'incidente alla centrale nucleare di Three Mile Island fu causato da un evento di questo genere. Il nocciolo dell'impianto HTGR e il sistema di generazione del vapore sono alloggiati in un recipiente in calcestruzzo precompresso le cui pareti hanno uno spessore di circa 4,5 metri.

L'unico HTGR negli Stati Uniti sorge a Denver, Colorado. È l'impianto di Fort St. Vrain, progettato dalla General Atomic, di proprietà della Public Service Company of Colorado e gestito da questa società. L'impianto, della potenza nominale di 330 MWe, è entrato in funzione tre anni fa e a tutt'oggi ha prodotto energia elettrica per oltre due miliardi di chilowattora. A varie riprese la circolazione forzata del fluido refrigerante nel nocciolo del reattore è stata interrotta per intervalli di tempo anche di 15 minuti, senza alcun danno al nocciolo o agli elementi di combustibile.

mento dei neutroni ed è in grado di assorbire il calore emesso dai prodotti di fissione dopo l'arresto della reazione nucleare a catena. In un reattore raffreddato a elio e moderato a grafite la reazione nucleare viene arrestata mediante l'inserimento di barre di controllo, analoghe a quelle degli altri reattori, oppure mediante l'iniezione di palline di boro, che «avvelenano» la reazione. Nei reattori raffreddati ad acqua, la perdita del refrigerante, che funge anche da moderatore, provoca l'arresto della reazione.

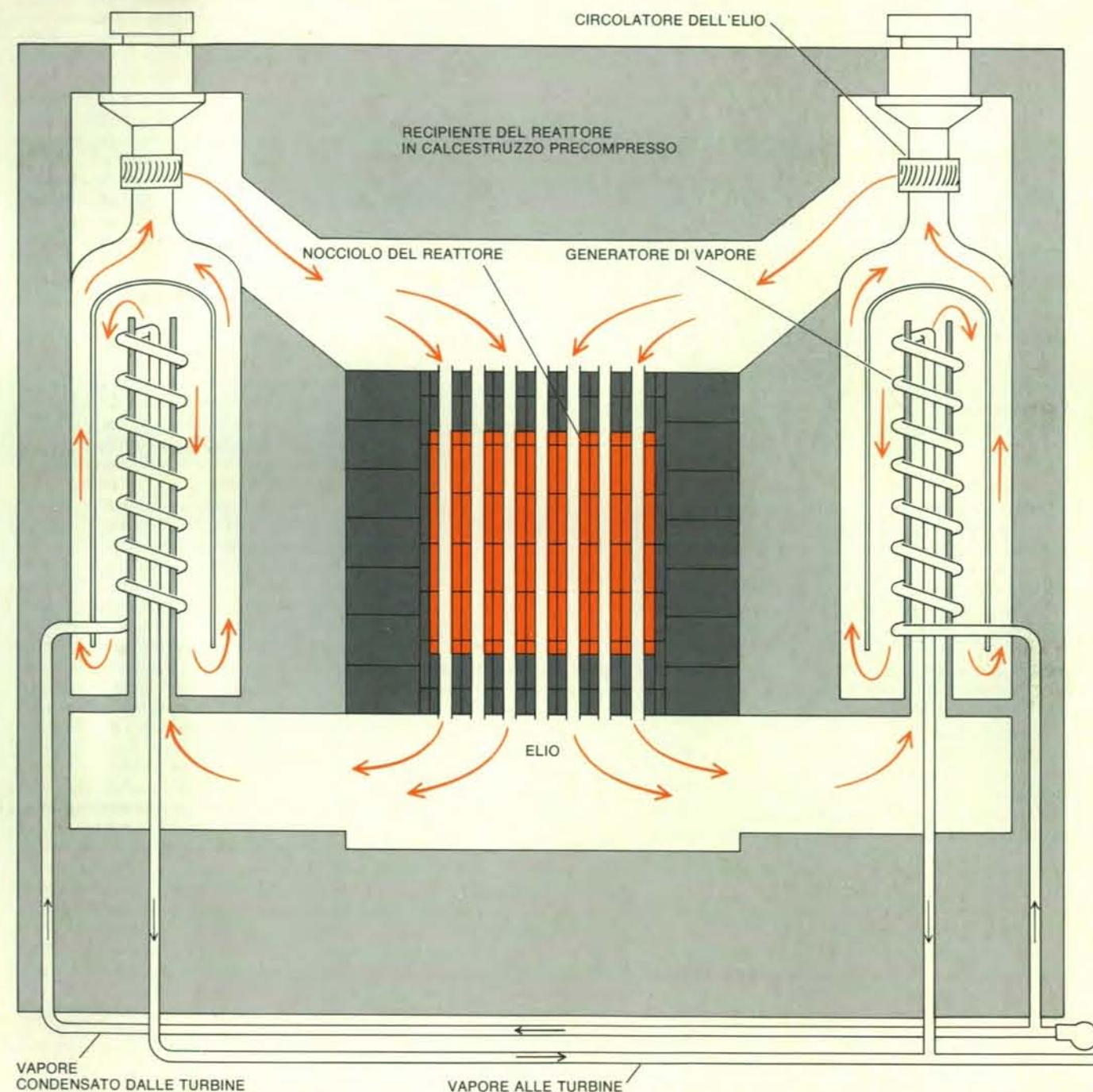
Se si guastassero i sistemi di raffreddamento di emergenza in un reattore ad acqua naturale, la temperatura del nocciolo crescerebbe anche dopo l'arresto della reazione perché i prodotti di fissione accumulatisi negli elementi di combustibile continuano a emettere energia a un tasso elevato. Al momento dell'arresto, il calore di decadimento corrisponde a circa il 7 per cento della potenza termica del reattore, cioè a circa 210 megawatt per un reattore ad acqua da 3000 megawatt termici (equivalenti a una potenza elettrica di 1000 MW). Si è valutato che in un incidente di perdita di refrigerante la temperatura dell'incamiciatura degli elementi di combustibile può raggiungere i 1650 gradi centigradi e che rotture del combustibile si possono verificare nello spazio di 50 secondi in un reattore ad acqua in pressione e di due minuti in un reattore ad acqua bollente. Per un reattore raffreddato a elio, in un incidente analogo comportante la depressurizzazione del sistema e il guasto completo del sistema di circolazione dell'elio, passerebbe più di un'ora prima che la temperatura del nocciolo possa raggiungere i 1650 gradi centigradi. A quella temperatura non si verificherebbero danni né alle particelle rivestite, né agli elementi di grafite del combustibile di un reattore a gas. Le particelle di combustibile e la grafite resistono egregiamente a temperature fino a 2200 gradi centigradi, temperature che sarebbero raggiunte non prima di 10 ore dall'inizio dell'incidente. In breve, il tempo a disposizione è più che sufficiente per l'adozione delle misure più appropriate per il ripristino del flusso di elio.

I pregi dei reattori a grafite raffreddati a gas hanno avuto un ampio riconoscimento al di fuori degli Stati Uniti. Negli anni cinquanta e sessanta, mentre gli Stati Uniti si rivolgevano ai reattori ad acqua naturale, Gran Bretagna e Francia sviluppavano reattori raffreddati a gas e moderati a grafite, il cui fluido refrigerante era anidride carbonica anziché elio. La Gran Bretagna ha attualmente in servizio o in costruzione oltre 40 reattori a gas, la Francia sette, e uno a testa Italia, Spagna e Giappone. L'esperienza di esercizio acquisita con i reattori a gas europei ammonta a oltre 600 anni-reattore. Tali reattori hanno contribuito per circa un quinto al totale dell'energia elettronucleare prodotta complessivamente a tutt'oggi nell'Europa occidentale, in Giappone e negli Stati Uniti.

Lo sviluppo dei reattori a gas francesi e britannici era a uno stadio preliminare nel 1956, quando un gruppo di fisici, formatisi prevalentemente al Los Alamos Scientific Laboratory, si riunì a La Jolla, in California, per affrontare il progetto di un reattore, che rispetto a quelli allora già sviluppati, fosse più efficiente e nello stesso tempo dotato di una maggiore flessibilità intrinseca. Tra i componenti del gruppo ricordiamo H. A. Bethe della Cornell University, Freeman J. Dyson dell'Institute for Advanced Study, Peter Fortescue dell'Atomic Energy Research Establishment di Harwell, in Inghilterra, e Frederic de Hoffmann, l'allora presidente della General Atomic. Iniziò così lo sviluppo del concetto di reattore ad alta temperatura raffreddato a gas, o HTGR (High Temperature Gas-cooled Reactor), cui contribuirono le conoscenze francesi e britanniche, sperimentato a Peach Bottom e, su scala maggiore, a Fort St. Vrain. Grazie alla grande disponibilità di elio negli Stati Uniti, quale fluido refrigerante venne scelto quel gas anziché l'anidride carbonica; rispetto a questo gas l'elio ha i notevoli vantaggi di essere stabile in presenza dell'alto flusso di radiazioni nel reattore, di non diventare radioattivo, di essere chimicamente inerte e di possedere eccellenti caratteristiche di trasmissione del calore.

I vantaggi degli HTGR sono stati illustrati da Joseph M. Hendrie, presidente della Nuclear Regulatory Commission, in una relazione al Congresso nel marzo 1980. Nella relazione egli afferma testualmente che tali reattori «hanno un rendimento confrontabile con quello dei migliori impianti termoelettrici a combustibile fossile e sono nettamente più efficienti dei reattori raffreddati ad acqua. Oltre a un maggiore rendimento termico gli HTGR hanno un migliore rendimento energetico, ossia consentono un migliore sfruttamento dell'uranio combustibile, superiore probabilmente del 15-20 per cento rispetto a quello ipotizzabile per i reattori ad acqua naturale». Inoltre, «gli HTGR presentano vantaggi dal punto di vista della sicurezza, essendo impianti nei quali non è indispensabile intervenire con grande urgenza se qualcosa non va come deve, in quanto la struttura del nocciolo è un grosso blocco di grafite, un materiale stabile e resistente alle alte temperature; ne consegue che se si verifica una caduta di potenza o un guasto ai circolatori dell'impianto gli operatori hanno tutto il tempo per riflettere con calma sulle azioni da intraprendere».

I primi reattori a gas realizzati in Gran Bretagna vennero denominati Magnox, perché le barre di combustibile, a uranio non arricchito, erano rivestite da una lega di magnesio. Il nocciolo del reattore, contenente varie tonnellate di grafite, era alloggiato in un recipiente a pressione voluminoso e costoso in acciaio, molto più grande di quelli adottati per i reattori ad acqua naturale. Nel 1958 i tecnici francesi dimostrarono che il recipiente in acciaio poteva essere sostituito da un recipiente in calcestruzzo precompresso, di dimensioni tali da potere alloggiare l'intero sistema del reattore, inclusi i generatori di vapore. Il recipiente a pressione in calcestruzzo precompresso, PCRV (Prestressed-Concrete Reactor Vessel), è costantemente mantenuto in compressione per mezzo di una rete di tensori di acciaio sovradimensionati in numero, controllabili, registrabili e sostituibili all'occasione. La tenuta rispetto a eventuali fughe di gas è assicurata da un involucro in acciaio fissato alla parete interna del PCRV; l'involucro funge solo quale membrana impervia per il contenimento del fluido refrigerante. L'involucro e le pareti del PCRV sono raffreddate con acqua che circola in tubazioni saldate sulla superficie esterna del recipiente.

In seguito i PCRV vennero adottati per tutti gli impianti a gas francesi e britannici. Grazie anche all'elevato livello di sicurezza ottenuto con l'adozione di questo tipo di recipiente, la Gran Bretagna decise di realizzare, in prossimità di siti urbani, impianti di seconda generazione denominati reattori a gas avanzati (AGR, Advanced Gas Reactors). Il combustibile di tali reattori era ossido di uranio, un materiale ceramico rivestito di acciaio inossidabile. L'innovazione era resa possibile dall'adozione di uranio leggermen-

Il nocciolo e i generatori di vapore dell'impianto HTGR sono alloggiati entro un massiccio recipiente a pressione in calcestruzzo precompresso (PCRV). Per un impianto da 860 MWe il PCRV ha un diametro di 31 metri e un'altezza di 25 metri. (Nella pagina successiva è rappresentato, alla stessa scala, un reattore ad acqua naturale di potenza poco superiore.) Il nocciolo di grafite di un reattore HTGR da 860 MWe occupa un volume cilindrico di 7,9 metri di diametro e di 6,4 metri di altezza. Quattro circolatori primari, di cui solo due sono rappresentati in questa sezione trasversale, fanno circolare l'elio alla pressione di 73,9 chilogrammi per centimetro quadrato in circa 27 000 canali verticali praticati entro il nocciolo. L'elio fuoriesce dal nocciolo del reattore alla temperatura di 685 gradi centigradi ed entra dal basso nei generatori di vapore, passando due volte in una serie di tubi elicoidali e diritti. L'acqua entra in ebollizione entro i tubi nel percorso ascendente e viene ulteriormente riscaldata nel percorso discendente, fuoriuscendo sotto forma di vapore surriscaldato alla temperatura di 538 gradi centigradi e una pressione di 176 chilogrammi per centimetro quadrato. Non sono rappresentati i tre circuiti di raffreddamento con scambiatori di calore raffreddati ad acqua, capaci di smaltire il calore dell'elio in circolazione nel caso di fuori servizio dei circuiti di generazione di vapore. Dopo un arresto del reattore i prodotti di fissione nel nocciolo liberano una quantità di calore inizialmente elevata che però diminuisce esponenzialmente.

te arricchito. Con il nuovo combustibile gli AGR erano in grado di funzionare a temperature superiori rispetto ai reattori Magnox e di «bruciare» una maggiore quantità dell'uranio 235 presente, prima che si rendesse necessario il ricambio del combustibile. Con l'aumento della temperatura il rendimento del processo di generazione di energia elettronucleare salì da circa il 30 per cento a poco più del 40 per cento.

Negli Stati Uniti, negli anni cinquanta e sessanta, l'AEC, l'Atomic Energy Commission (predecessore dell'attuale Department of Energy), concretizzò il proprio interesse per i reattori raffreddati a gas finanziando gli studi di diversi concetti di reattori avanzati. Uno dei principali obiettivi dell'AEC era quello di ridurre la quantità di uranio necessaria per unità di energia elettrica prodotta; a quel tempo le risorse di uranio sembravano insufficienti rispetto ai prevedibili fabbisogni. Da quegli studi venne una chiara indicazione in favore di reattori autofertilizzanti o di reattori convertitori avanzati. In un autofertilizzante, si forma almeno un atomo di nuovo combustibile per ogni atomo di combustibile consumato. In un convertitore avanzato, per ogni atomo di combustibile consumato si formano generalmente da 0,7 a 1 atomi di nuovo combustibile. Nei reattori ad acqua, invece, per ogni atomo di combustibile consumato si formano da 0,5 a 0,6 atomi di nuovo combustibile. Il reattore ad alta temperatura raffreddato a gas e moderato a grafite è un convertitore avanzato, ed è uno dei concetti di reattore emersi dalla inevitabile selezione condotta su quelli studiati nell'ambito dell'iniziativa dell'AEC. Le industrie produttrici di elettricità si sono espresse molto positivamente sul reattore HTGR, sia perché rispetto al reattore ad acqua naturale è competitivo in termini di costi di capitale, sia perché consente l'adozione del ciclo combustibile uranio-torio, con un basso consumo di uranio e quindi bassi costi di combustibile.

Nel successivo stadio di sviluppo dei reattori a gas sono confluiti almeno due degli importanti aspetti tecnologici ideati separatamente in Europa e negli Stati Uniti: l'elio si sostituisce all'anidride carbonica quale fluido refrigerante e il nocciolo del reattore viene caricato con combustibile che consente di fare a meno dell'incamiciatura metallica. Queste due caratteristiche sono presenti non solo a Peach Bottom e a Fort St. Vrain, ma anche in due reattori europei: un reattore sperimentale a elio da 20 MW termici, in funzione dal 1965 al 1976 nell'Inghilterra meridionale, e un reattore HTGR da 15 MWe (denominato AVR), in esercizio commerciale in Germania dal 1967, con temperatura di uscita del gas fino a 950 gradi centigradi. (La temperatura dell'acqua in uscita dal nocciolo di un reattore ad acqua naturale in pressione è di circa 321 gradi centigradi.) In Germania è attualmente in costruzione un impianto da 300 MWe, basato sul concetto AVR, la cui entrata in funzione è prevista per il 1984 o 1985. Negli Stati Uniti, il reattore di Fort St. Vrain da 330 MWe ha prodotto energia elettrica per più di due miliardi di chilowattora dal 1978, fornendo una dimostrazione delle prestazioni del combustibile e delle caratteristiche di sicurezza di un moderno progetto HTGR. Il reattore è stato sottoposto a prove in condizioni di emergenza, inclusa la prova di totale perdita della circolazione forzata del fluido refrigerante, senza che si dovesse registrare alcun danneggiamento del nocciolo o di altri componenti principali dell'impianto.

Sulla base dell'esperienza di Fort St. Vrain, la General Atomic, in collaborazione con la Gas-Cooled Reactor Associates (un consorzio di società statunitensi produttrici di elettricità) e con il Department of Energy, ha sviluppato il progetto di riferimento di un impianto HTGR da 860 MWe. Si tratta di un progetto semplificato e cautelativo, particolarmente curato dal punto di vista sia della sicurezza sia della protezione da rischi del capitale investito. Il nocciolo del reattore è contenuto in un recipiente a pressione in calcestruzzo precompresso a più cavità. L'elio esce dal nocciolo a 685 gradi centigradi (nel reattore di Fort St. Vrain a 812 gradi centigradi) e passa in quattro circuiti di raffreddamento primari, dove si genera vapore alla temperatura di 538 gradi centigradi e alla pressione di 176 chilogrammi per centimetro quadrato.

Il flusso di elio viene forzato in ogni circuito di raffreddamento da un circolatore azionato da un motore elettrico. (I circolatori di Fort St. Vrain sono invece azionati dal vapore.) Inoltre, il nocciolo è equipaggiato con un sistema di raffreddamento ausiliario costituito da tre circuiti, ciascuno in grado di erogare da solo la totalità del raffreddamento richiesto quando l'elio nel recipiente del reattore è alla normale pressione di esercizio di 73,9 chilogrammi per centimetro quadrato, oppure al 50 per cento del raffreddamento richiesto, se il recipiente è depressurizzato. L'elio che circola nel sistema di raffreddamento ausiliario viene raffreddato con l'acqua messa in circolazione mediante elettropompe, che possono essere alimentate, all'occorrenza, da generatori diesel.

L'accoppiamento di un gas stabile e inerte quale fluido refrigerante con una struttura del nocciolo in grafite, resistente alle alte temperature, consente la produzione di vapore alle condizioni di temperatura e pressione corrispondenti a quelle dei moderni impianti termoelettrici a combustibile fossile. Il rendimento netto previsto dal progetto di riferimento di tale HTGR ammonta al 38,5 per cento, di poco inferiore al 39,2 per cento ottenuto a Fort St. Vrain. La lieve riduzione dipende dalla maggiore semplificazione del sistema di generazione di vapore e dall'introduzione di ulteriori margini operativi e di sicurezza.

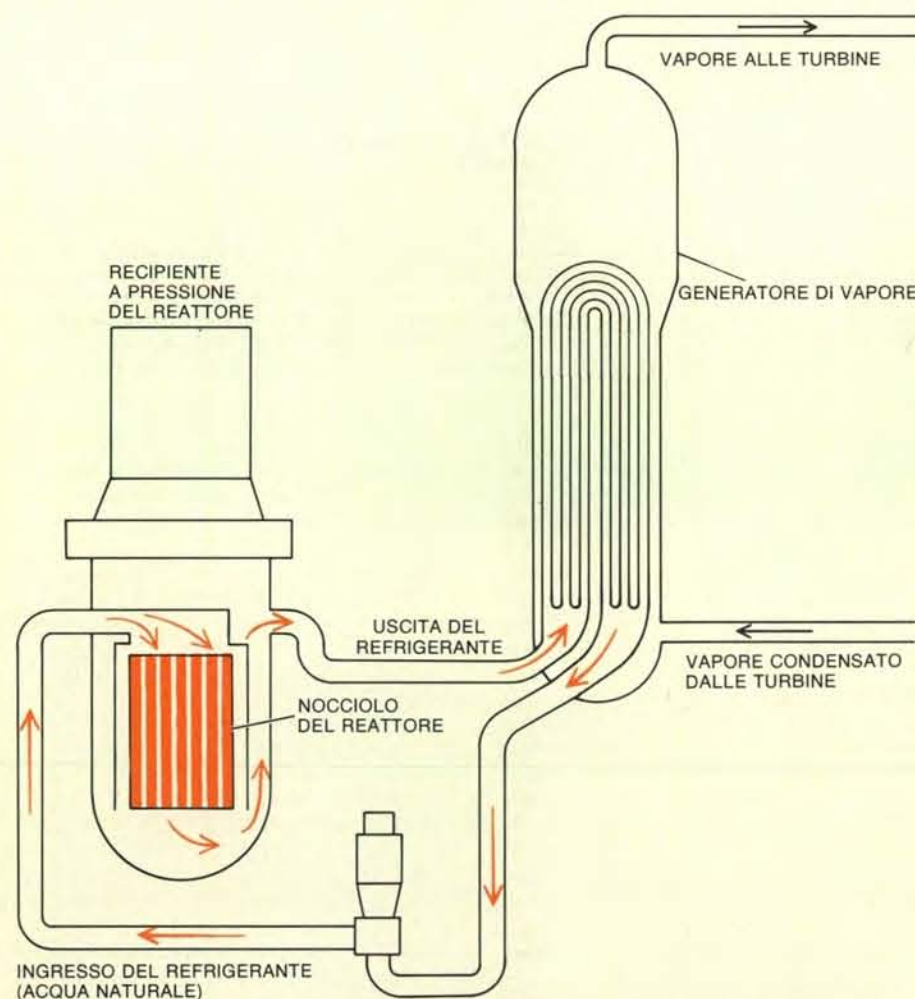
Il gas elio, refrigerante che all'interno del sistema non può venire condensato in forma liquida, ha la proprietà fondamentale di seguire una relazione lineare tra temperatura e pressione; ne consegue che i rilevamenti strumentali di temperatura e di pressione possono fornire il controllo indipendente delle due grandezze. Non essendovi interfacce gas-liquido, come nei reattori ad acqua bollente (e nei reattori ad acqua in pressione in certe condizioni di emergenza), un solo preciso segnale - quello di pressione - indica sempre la presenza e lo stato fisico del fluido refrigerante. Una rapida depressurizzazione del sistema di raffreddamento primario può verificarsi senza che si debba paventare la formazione di vuoti o lo scoprimento parziale del nocciolo, eventi che invece possono verificarsi in seguito alla liberazione di pressione da parte di acqua che si trovi al di sopra del proprio punto di ebollizione atmosferico.

L'esperienza di Fort St. Vrain ha permesso di verificare alcuni vantaggi degli HTGR dal punto di vista operativo e della sicurezza. Il personale addetto all'esercizio e alla manutenzione è stato esposto a dosi di radiazioni ben al di sotto dei limiti consentiti per il personale degli impianti nucleari. Su un totale di diverse centinaia di addetti, solo per meno di una decina è stato possibile rilevare dosi misurabili di radiazioni.

L'impianto di Fort St. Vrain ha reagito senza scosse e senza affanno alle variazioni di carico provocate sia da escursioni transitorie nel ciclo di generazione di potenza sia a temporanei fuori servizio di componenti o sistemi. Il nocciolo degli HTGR è di grosse dimensioni e per unità di volume libera una minore quantità di calore rispetto ai reattori ad acqua naturale; inoltre la gran massa del nocciolo, che contiene circa 1500 tonnellate di grafite, ha un'elevata capacità di assorbimento di calore nel caso che il flusso di raffreddamento venga a ridursi o a interrompersi. Per questi motivi il reattore risponde lentamente a un imprevisto guasto d'esercizio, consentendo agli operatori un ampio margine di tempo (ore in luogo di minuti) per attuare i provvedimenti più opportuni.

A Fort St. Vrain cinque guasti di questo genere hanno provocato l'interruzione della circolazione forzata dell'elio per lunghi periodi di tempo, senza provocare rilevabili incrementi della temperatura del nocciolo né causare rischi di alcun genere per l'impianto o il combustibile. Viene così praticamente eliminato il rischio di danneggiamento del reattore o del nocciolo in seguito a errori degli operatori. Grazie alla stabilità termica degli HTGR risulta semplificato il progetto del sistema di arresto del reattore mediante inserimento di assorbitori neutronici e del sistema di raffreddamento d'emergenza. Se la normativa lo consente, c'è anche tempo più che sufficiente per l'azionamento manuale di tali sistemi. In seguito all'incidente di Three Mile Island, la Nuclear Regulatory Commission ora prescrive la presenza a pieno tempo sul sito degli impianti nucleari di un esperto, denominato consulente tecnico d'esercizio. L'unico impianto che deroga a questa norma è quello di Fort St. Vrain, per il quale non è necessario che l'esperto sia presente in centrale, ma è sufficiente che si renda reperibile entro un'ora dalla eventuale chiamata.

Il recipiente a pressione in calcestruzzo precompresso rappresenta uno dei principali aspetti di sicurezza distintivi del progetto di questo reattore. In primo luogo, una rottura catastrofica del PCRV è un evento che l'analisi dei rischi definisce incredibile, tanto remota è la probabilità che possa verificarsi. Ciascun tensore in acciaio che contribuisce alla resistenza del PCRV è indipendente e ridondante e il recipiente si trova con continuità in uno stato di compressione. In secondo luogo, il recipiente è progettato per resistere a una pressione massima più che doppia di quella normale d'esercizio, ossia a 170 chilogrammi per centimetro quadrato. Qualsiasi fessurazione che dovesse risultare eventualmente da un eccesso di pressione non provocherebbe altro che una lenta perdita di gas, che tenderebbe a cessare spontaneamente con una leggera riduzione della pressione. In terzo luogo, la depressurizzazione completa può verificarsi solo in caso di rottura nella zona di inserimento delle tubazioni o delle piccole linee di servizio che attraversano la parete del PCRV. Rotture di questo tipo rappresentano un evento estremamente

VAPORE ALLE TURBINE
RECIPIENTE A PRESSIONE DEL REATTORE
GENERATORE DI VAPORE
USCITA DEL REFRIGERANTE
VAPORE CONDENSATO DALLE TURBINE
NOCCIOLO DEL REATTORE
INGRESSO DEL REFRIGERANTE (ACQUA NATURALE)

L'impianto nucleare ad acqua naturale in pressione è il più diffuso negli Stati Uniti, con 44 unità in esercizio. Altri 24 reattori sono del tipo ad acqua bollente, in cui il calore del nocciolo viene trasmesso dal vapore anziché dall'acqua pressurizzata. Questa illustrazione rappresenta il nocciolo di un reattore ad acqua naturale in pressione da 1100 MWe. Il nocciolo è alloggiato in un recipiente a pressione in acciaio di circa 4,5 metri di diametro, 12,2 metri di altezza e con uno di spessore da 15 a 27,5 centimetri; il recipiente è progettato per una pressione interna di esercizio di 156 chilogrammi per centimetro quadrato. L'acqua di raffreddamento esce a 321 gradi centigradi e passa in quattro circuiti generatori di vapore, di cui uno solo è qui raffigurato. All'uscita dai generatori il vapore ha una temperatura di 282 gradi centigradi e una pressione di 70 chilogrammi per centimetro quadrato. A queste condizioni di temperatura e pressione il rendimento termico dell'impianto è solo del 32-33 per cento, mentre è del 38,5 per cento quello di un impianto HTGR.

Sono qui raffigurati, rispettivamente a sinistra e a destra, una barra di combustibile e un blocco di combustibile di un reattore HTGR. La barra, lunga circa 6,25 centimetri, è costituita da decine di migliaia di particelle combustibili inglobate in una matrice di grafite. Ogni blocco di combustibile, che misura circa 35 centimetri di larghezza e 77,5 centimetri di altezza, è costituito da 1656 barre combustibili compattate in una struttura esagonale. I numerosi canali praticati nel blocco servono per il passaggio del flusso di elio. Il grosso foro centrale alloggia il meccanismo per l'inserimento dei blocchi di combustibile nel nocciolo del reattore. Il nocciolo di un reattore da 860 MWe è costituito da 3512 blocchi. Ogni blocco combustibile, del peso di circa 122 chilogrammi, contiene in media 0,7 chilogrammi di uranio 235 e 15,9 chilogrammi di torio 232. In quattro anni di permanenza nel nocciolo un blocco di combustibile fornisce una quantità di energia equivalente a quella ottenibile da 2500 tonnellate di carbone o da 12 000 barili di olio combustibile. Se venisse recuperato e si riciclasse l'U-235 e l'U-233 non utilizzati, l'energia equivalente ottenibile dal combustibile originario salirebbe a circa 11 000 tonnellate di carbone e a 54 000 barili di olio combustibile.

Confronto tra le caratteristiche di esercizio di un reattore HTGR da 860 MWe *(in colore)* e quelle di un reattore ad acqua in pressione della stessa capacità *(in grigio)*. Il minor consumo di combustibile in un impianto HTGR può essere attribuito in parte al più elevato rendimento termico e in parte al fatto che per ogni atomo di uranio consumato nello HTGR si producono circa 0,7 atomi di combustibile nuovo. In un reattore ad acqua in pressione il numero di nuovi atomi prodotti per ogni atomo consumato è inferiore a 0,5. Con un ciclo del combustibile senza recupero, in entrambi i tipi di impianto una certa frazione di atomi di U-238 o di Th-232 viene trasformata in isotopi di plutonio o uranio, che in parte vengono utilizzati prima del ricambio del combustibile. Se fosse consentito il riciclo del combustibile esaurito (non ammesso dalle direttive del precedente governo degli Stati Uniti), il combustibile HTGR dovrebbe essere una miscela di uranio ad alto arricchimento (circa 93 per cento in U-235) e di torio. Parte del torio sarebbe trasformata in U-233 fissile, che verrebbe recuperato e riciclato per sostituire l'U-235 nelle successive ricariche.

improbabile. Inoltre, in corrispondenza di ciascun punto di inserimento, il nocciolo è dotato di limitatori di flusso, per impedire la liberazione rapida del gas che potrebbe causare danni strutturali al nocciolo o al sistema di raffreddamento.

Rispetto alla generazione attuale di reattori le superiori prestazioni degli HTGR avvantaggiano questi impianti anche dal punto di vista ambientale. Poiché il rendimento degli HTGR è circa del 39 per cento, mentre quello dei reattori ad acqua naturale è del 33 per cento, gli HTGR liberano nell'ambiente circostante una quantità di calore di scarto inferiore del 25 per cento. Ne consegue che viene ridotto in proporzione il rischio di danni all'ecosistema acquatico. Nel caso di smaltimento del calore mediante scarico dell'acqua calda in un lago o in un fiume nelle vicinanze dell'impianto, se per lo smaltimento del calore vengono adottate torri di raffreddamento, queste consumano minori quantità di acqua, oltre a essere più ridotte in dimensioni e meno costose. Se si usano bacini di raffreddamento, a parità di estensione del bacino, è possibile installare un impianto HTGR di potenza superiore di un terzo rispetto a un impianto ad acqua naturale, senza che vengano superati i limiti di temperatura autorizzati per il bacino. Se si adottano torri di raffreddamento a secco, nei casi prescritti dalla normativa ambientale o nei casi in cui l'acqua disponibile sia scarsa, la riduzione di potenza a temperatura esterna elevata risulterebbe per un impianto HTGR metà di quella corrispondente per gli impianti nucleari attualmente in esercizio. Conseguentemente, un impianto HTGR può sorgere in siti remoti aridi o semiaridi con una penalizzazione minore in termini di costi.

Il livello di radioattività nelle emissioni normali degli impianti nucleari viene accuratamente controllato. Rispetto ai reattori ad acqua naturale gli HTGR emettono, nei vari processi che hanno luogo nell'impianto, quantità inferiori di radioattività e a concentrazioni più basse. Inoltre, gli HTGR sono equipaggiati con sistemi che assicurano che le emissioni di radioattività dall'impianto all'ambiente siano praticamente nulle. Le operazioni di manutenzione di routine possono produrre piccole quantità di rifiuti liquidi di bassa attività (meno di 7500 litri all'anno, con un'attività totale associata inferiore a 150 curie), trasportabili con facilità lontano dal sito o ivi immagazzinabili. I rifiuti solidi prodotti da un impianto HTGR non superano i 56 metri cubi all'anno. Circa l'80 per cento è rappresentato da rifiuti di basso livello, ossia solo lievemente contaminanti (quali carta, parti di filtri e resine esaurite), facilmente allontanabili dal sito in serbatoi da interrare o da bruciare praticamente senza effetti sull'ambiente. Il rimanente 20 per cento è rappresentato da rifiuti di livello intermedio, essenzialmente costituiti dai blocchi del riflettore che devono venire sostituiti periodicamente. Questi rifiuti si possono trasportare lontano dal sito in appositi contenitori schermati di circa 210 litri di capacità, per l'immagazzinamento sicuro a lungo termine. I sistemi di purificazione dell'elio e di recupero gassoso previsti negli impianti HTGR standard riducono i livelli della radioattività presente nei gas liberati a valori di diversi ordini di grandezza inferiori ai cinque millirem annui previsti dalla normativa ufficiale. Il trizio (l'isotopo radioattivo dell'idrogeno) che si genera entro il sistema primario di un impianto HTGR viene rimosso nel sistema di purificazione dell'elio da un ossidante che lo converte in acqua triziata, la quale successivamente viene solidificata e trattata come rifiuto solido agevolmente isolabile non appena il trizio è decaduto. (La vita media del trizio è di 12,26 anni.)

Negli HTGR sono stati attuati vari accorgimenti che semplificano l'esercizio, la manutenzione e la ricarica. Ad esempio, tutto il fluido refrigerante primario, l'elio, è confinato all'interno del recipiente a pressione in calcestruzzo precompresso. Lo stesso PCRV costituisce la schermatura indispensabile per il personale, in modo che i lavori di manutenzione si possono svolgere in tutti i punti dell'edificio del reattore con l'impianto in funzione. Poiché nell'intero sistema secondario di generazione del vapore non vi è praticamente traccia di radioattività, l'esercizio e la manutenzione di tutta la parte dell'impianto relativa al ciclo del vapore esterna al PCRV, incluso il sistema turbogeneratore, possono avvenire esattamente come in un impianto termoelettrico a combustibile fossile. Inoltre, tutti i componenti e i sistemi connessi con il ciclo di vapore e dell'acqua di alimentazione di un impianto HTGR sono di dimensioni ridotte rispetto a un impianto ad acqua naturale e quindi di più agevole manutenzione, essendo i flussi di vapore nelle turbine di un impianto HTGR pari al 60 per cento del corrispondente flusso di vapore in un impianto ad acqua naturale. In generale, in un impianto HTGR i costi di manutenzione, riparazione e gestione sono inferiori rispetto a un impianto ad acqua naturale, dal momento che l'elio, a differenza dell'acqua, è inerte, non radioattivo e non corrosivo.

Uno dei grandi vantaggi del gas quale fluido refrigerante è dato dalla sua trasparenza, che rende possibile l'ispezione visiva di varie zone entro il PCRV. Le proprietà di schermatura offerte dal progetto del PCRV rendono possibile l'effettuazione di varie operazioni di ispezione e manutenzione con reattore in funzione, il che riduce i periodi in cui il reattore deve essere arrestato per questi scopi.

Praticamente tutti i componenti strutturali del PCRV, quali i tensori verticali e gli avvolgitori dei cavi di precompressione circonferenziale, possono essere ispezionati visivamente con il reattore in funzione. Taluni componenti particolari vengono tenuti sotto sorveglianza continua per il rilevamento di variazioni nella tensione o della presenza di sforzi, che indicherebbero un peggioramento delle prestazioni dell'impianto. Se necessario, tutti i componenti strutturali possono essere sostituiti. Tutte le superfici esterne in calcestruzzo, a eccezione di quelle nelle immediate vicinanze dei punti di inserimento delle barre di controllo, possono essere ispezionate visivamente con reattore in funzione. I punti di inserimento delle barre di controllo nel PCRV e le superfici in prossimità di tali punti di inserimento possono essere agevolmente ispezionate nel corso della ricarica.

La sicurezza intrinseca degli HTGR è rappresentata in questo grafico, che riporta il confronto tra l'andamento della temperatura del nocciolo di un reattore HTGR, di un reattore ad acqua in pressione e di un reattore ad acqua bollente, in seguito a un ipotetico incidente di perdita di refrigerante o di arresto della circolazione forzata. Nei reattori raffreddati ad acqua, la perdita di acqua, che funge anche da moderatore, provoca l'arresto automatico della reazione nucleare. Negli HTGR la reazione deve essere arrestata inserendo le barre di controllo che assorbono i neutroni. Al momento dell'arresto, al decadimento dei prodotti di fissione nel combustibile corrisponde una produzione di calore pari al 7 per cento della potenza termica del reattore. Questo valore scende all'1 per cento entro due ore circa, e allo 0,5 per cento entro 24 ore. Nei reattori raffreddati ad acqua, in assenza di raffreddamento di emergenza, la temperatura dell'incamiciatura del combustibile sale in meno di due minuti a 1645 gradi centigradi, provocandone la rottura. Negli HTGR la massa del moderatore di grafite assorbirebbe il calore emesso dai prodotti di fissione, in modo che i 1645 gradi non verrebbero raggiunti se non dopo un'ora. Solo dopo almeno dieci ore di assenza del raffreddamento forzato si raggiungerebbe una temperatura sufficientemente elevata da provocare il danneggiamento del nocciolo di grafite (circa 2200 gradi centigradi).

ANNO	DISPONIBILITÀ DEL SISTEMA NUCLEARE DI GENERAZIONE DEL VAPORE (PER CENTO)	DISPONIBILITÀ DELL'IMPIANTO (PER CENTO)	FATTORE DI UTILIZZAZIONE GLOBALE DELL'IMPIANTO (PER CENTO)
1967	81	78	69
1968	88	88	82
1969	86	84	67
1970	95	95	88
1971	90	87	78
1972	71	71	58
1973	95	94	78
1974	96	95	70
MEDIA 1° NOCCIOLO	85	83	73
MEDIA 2° NOCCIOLO	89	88	74
MEDIA SULLA VITA TOTALE DELL'IMPIANTO	88	86	74

La statistica qui riportata dimostra l'affidabilità del primo impianto HTGR progettato dalla General Atomic, quello di Peach Bottom. Se si escludono i fuori servizio programmati o dovuti a motivi indipendenti dal sistema nucleare, la disponibilità dell'impianto è stata dell'88 per cento. Il fattore di utilizzazione nei sette anni e mezzo di esercizio dell'impianto è stato del 74 per cento, superiore quindi al 66 per cento tipico degli impianti ad acqua naturale in servizio commerciale.

Le particelle di combustibile messe a punto per i reattori HTGR misurano 0,075 centimetri di diametro, circa le dimensioni di un granello di sabbia. In questa figura una sezione equatoriale di una particella è ingrandita 150 volte. Il combustibile nucleare è il materiale cristallino visibile al centro. Esso consiste di ossicarburo di uranio, con un arricchimento del 93 per cento nell'isotopo fissile uranio 235, valore che consente le più elevate prestazioni del combustibile. Gli strati di carbonio e di carburo di silicio vengono depositati mediante un processo ad alta temperatura.

La recente esperienza delle ricariche di Fort St. Vrain ha messo in evidenza la convenienza per gli HTGR degli elementi di combustibile a blocchi. Circa 240 elementi, pari a un sesto del totale, sono stati estratti dal nocciolo e sostituiti con combustibile fresco; gli altri 1240 elementi non sono stati spostati. Per tali operazioni l'esposizione alle radiazioni degli addetti è stata così ridotta che per misurarla si è ricorsi a un rivelatore di microrem. Estrapolando dai dati disponibili si può calcolare che la somma delle esposizioni integrate per l'intera operazione di ricarica successiva al funzionamento del reattore a piena potenza ammonterebbe a meno di cinque rem-uomo. L'attuale normativa federale prevede il limite di cinque rem nel periodo di un anno.

L'elemento di combustibile HTGR è un blocco di grafite, esagonale in sezione trasversale, di 35 centimetri di larghezza e di 77,5 centimetri di altezza. Il blocco è forato longitudinalmente in 72 canali per il fluido refrigerante e in 138 canali ciechi per il combustibile. La grafite è ideale come moderatore e come materiale strutturale, in quanto la sua resistenza cresce con la temperatura. Nel progetto di riferimento i blocchi di grafite sono sovrapposti in pile di otto. Questa sistemazione a segmentazione assiale agevola la fabbricazione e le operazioni di manutenzione e di ricarica.

Questa configurazione a blocchi così vantaggiosa è stata resa possibile dallo sviluppo delle particelle di combustibile con rivestimento speciale. Il nucleo di ogni particella è costituito da una microsfera di ossicarburo di uranio (convenientemente arricchita in uranio 235) di circa 0,025 centimetri di diametro. Attorno a ogni nucleo vengono depositati ad alta temperatura sottili strati di carbonio, carbonio pirolitico e carburo di silicio, in modo da formare una sferula a rivestimento ermetico, del diametro di circa 0,075 centimetri. Un'analoga procedura costruttiva si adotta per le particelle di torio. Questa tecnica assicura il contenimento dei prodotti di fissione. Le sferule vengono sottoposte a gruppi di 2000 alle prove di integrità strutturale, nel corso delle quali sono esposte a flussi di radiazione che simulano l'ambiente interno del reattore. Il procedimento di fabbricazione delle particelle, semiautomatico, è accoppiato a una rigorosa procedura di verifica che garantisce un accurato controllo di qualità.

Se in seguito a imprevedibili condizioni operative si dovesse verificare la rottura del rivestimento delle particelle di combustibile, con conseguente emissione di prodotti di fissione, la rottura sarebbe circoscritta alla sola zona del nocciolo direttamente interessata. Per la maggior parte dei reattori, invece, in cui l'incamiciatura riveste il combustibile per tutta la lunghezza del nocciolo, un evento incidentale che provochi la rottura di una piccola parte di tale rivestimento può causare l'emissione dei prodotti di fissione dell'intera barra di combustibile. Gli elementi di combustibile di Fort St. Vrain hanno dato prestazioni perfettamente corrispondenti a quelle previste dal progetto. Anzi, l'emissione di prodotti di fissione è stato ben al di sotto dei livelli previsti. In sostanza, nel combustibile HTGR si è dimostrato l'alto grado di affidabilità delle barriere contro l'emissione dei prodotti di fissione.

Le caratteristiche degli HTGR consentono l'adozione di un'ampia varietà di cicli per il combustibile nucleare. Il ciclo finora più studiato a fondo è il ciclo uranio-torio, in cui l'uranio ad altissimo arricchimento (93 per cento in U-235) funge da materiale fissile e il torio (Th-232) da materiale «fertile». Nel reattore il torio assorbe neutroni e viene alla fine trasformato nell'isotopo fissile uranio 233, che può essere riciclato nelle successive cariche di combustibile. Il combustibile del reattore di Fort St. Vrain è costituito di uranio arricchito al 93,5 per cento in U-235, in combinazione con torio. Il progetto dell'impianto consente l'adozione sia di uranio ad alto arricchimento sia di uranio a medio arricchimento (intorno al 20 per cento in U-235). I costi connessi al ciclo del combustibile HTGR sono praticamente equivalenti a quelli relativi agli altri reattori commerciali, a causa degli attuali vincoli sul ritrattamento del combustibile e sul riciclo. A meno di un mutamento nelle direttive del governo americano in questa materia, il combustibile esaurito viene immagazzinato a tempo indeterminato, senza possibilità di recupero dell'U-235 non esaurito, dell'U-233 o del plutonio formatisi nel corso dell'esercizio del reattore. Questo ciclo di combustibile è noto come «ciclo aperto».

Se si adottasse un ciclo uranio-torio di questo tipo con uranio ad alto arricchimento, un impianto HTGR nei suoi 40 anni di vita consumerebbe una quantità di uranio inferiore del 20 per cento rispetto a un reattore ad acqua naturale. Se per entrambi i tipi di impianto si potesse adottare il completo recupero dell'uranio e del plutonio, gli HTGR consumerebbero circa il 50 per cento in meno di uranio. Gli HTGR offrono quindi la possibilità di risparmiare una cospicua quantità di uranio sia con il ciclo aperto sia con il riciclo totale, purché il reattore sia progettato in modo da consentire l'adozione di uranio ad alto arricchimento. Si può apprezzare la portata di questo potenziale risparmio di uranio se si considera che il costo complessivo del combustibile riferito all'intera vita dell'impianto è confrontabile con il costo iniziale complessivo dell'impianto.

Negli ultimi sei anni sono stati annullati gli ordini per 55 impianti nucleari di potenza. Solo sei anni fa le società produttrici di elettricità statunitensi hanno espresso interesse per la realizzazione di 10 impianti HTGR. Quando il reattore di Fort St. Vrain funzionerà a piena potenza, il che è previsto per questa estate, e avrà dato dimostrazione della eccezionale sicurezza e affidabilità di cui sono certi i progettisti di questo impianto, è probabile che le società produttrici statunitensi guardino con favore agli HTGR nel caso di ripresa delle ordinazioni di impianti nucleari di potenza.

Gli archibatteri

Questi insoliti batteri non sono, come genealogia, né procarioti, né eucarioti. Questa scoperta significa che non esistono due, bensì tre linee di discendenza: gli archibatteri, i batteri veri e gli eucarioti

di Carl R. Woese

I primi naturalisti sostenevano che la vita sulla Terra presenta una fondamentale dicotomia: tutti gli esseri viventi si ripartiscono fra piante e animali. Quando vennero scoperti i microrganismi, vennero anch'essi suddivisi nello stesso modo. Quelli grossi e dotati di motilità furono considerati animali, quelli che chiaramente non si muovevano, e che comprendevano i batteri, furono inclusi tra le piante. A mano a mano che progredirono le conoscenze del mondo microscopico, fu evidente che una semplice classificazione in due tronconi non era più sufficiente. Vennero allora introdotte nuove categorie: i funghi, i protozoi e i batteri. Per finire, però, si ebbe una nuova semplificazione: sembrò che, dopo tutto, la vita potesse essere dicotomica, anche se a un livello più profondo, in particolare nella struttura della cellula vivente. Tutte le cellule parvero rientrare nell'una o nell'altra di due categorie: gli eucarioti, con cellule dotate di un nucleo ben formato, e i procarioti, che non presentano un nucleo del genere. Le piante e gli animali pluricellulari sono eucarioti, così come lo sono molti organismi unicellulari. Unici procarioti sono i batteri (comprendenti anche i cianobatteri, un tempo chiamati alghe azzurre).

Negli ultimi anni i miei colleghi e io siamo stati indotti a proporre una fondamentale revisione di questo quadro. Tra i batteri abbiamo trovato un gruppo di organismi che non sembra appartenere né all'una né all'altra delle due categorie fondamentali. Gli organismi che costituiscono l'oggetto del nostro studio sono procarioti nel senso che non hanno un nucleo e, difatti, visti dall'esterno, sembrano proprio batteri normali. Tuttavia, sotto l'aspetto biochimico e per la struttura di certe loro macromolecole, sono diversi dagli altri procarioti così come lo sono dagli eucarioti. Da un punto di vista filogenetico, non appartengono, dunque, né all'uno né all'altro gruppo, ma costituiscono un nuovo «regno primario», con una posizione completamente diversa nella storia e nell'ordine naturale della vita.

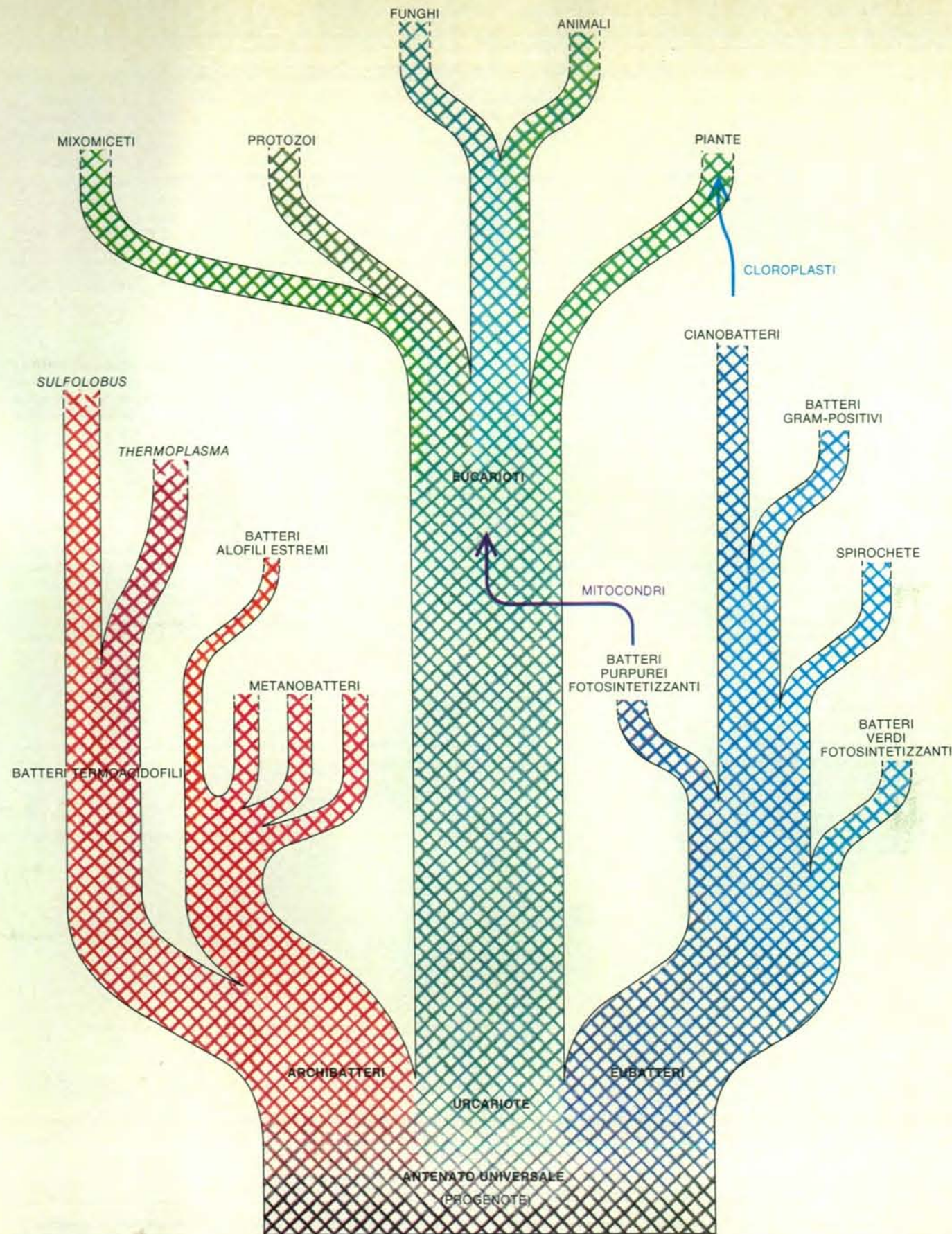
Abbiamo denominato questi organismi «archibatteri», termine che riflette un'ipotesi non comprovata circa la loro condizione evolutiva. Le prove filogenetiche fanno pensare che essi siano perlomeno altrettanto antichi degli altri gruppi principali. Inoltre, alcuni hanno una forma di metabolismo che sembra particolarmente ben adattata alle condizioni che si ritiene fossero prevalenti agli albori della storia della vita sulla Terra. Pertanto, è possibile che il gruppo di organismi di più recente definizione sia in realtà il più vecchio.

La documentazione evoluzionistica

L'età della Terra è di quattro miliardi e mezzo di anni; basandosi sui reperti fossili macroscopici, sembrerebbe che il pianeta sia stato abitato per meno di un settimo di questo tempo: l'intero corso evolutivo, dalle più antiche forme marine all'uomo, spazia per solo 600 milioni di anni. Invece le impronte fossili degli organismi cellulari, troppo piccole per essere viste a occhio nudo, raccontano una storia diversa. In particolare, i microfossili di batteri sono abbondanti nei sedimenti di ogni età: sono stati trovati nelle più antiche rocce sedimentarie intatte che siano note, i depositi australiani che risalgono a 3 miliardi e mezzo di anni fa. In un enorme lasso di tempo, durante il quale non esistevano ancora forme di vita superiori, i batteri fecero la loro comparsa e si irradiarono formando un'ampia varietà di tipi, presenti in un numero molto elevato di nicchie ecologiche. Quest'età dei microrganismi costituisce il periodo più importante nella storia della evoluzione, non solo per la sua durata, ma anche per la natura degli eventi evolutivi che si verificarono in quei miliardi di anni.

Fino a non molto tempo fa, tuttavia, non si sapeva quasi nulla di questo periodo. I microfossili batterici sono strutture che non danno grandi informazioni: ben poco si può dedurre, infatti, da una impronta lasciata da una sferetta o da un bastoncino. Le principali indicazioni paleontologiche sulla natura dei primi batteri non sono giunte da singoli microfossili, ma da strutture macroscopiche, dette stromatoliti, che si ritiene siano intrecci di filamenti batterici fossili, ossia colonie batteriche mineralizzate. Oggi, strutture di questo tipo sono formate in primo luogo da parecchi tipi di batteri con attività fotosintetica, in genere cianobatteri. Le stromatoliti di recente fossilizzazione assomigliano a quelle antiche in misura tale, che sembra assolutamente ragionevole pensare che anche le antiche strutture siano state prodotte da batteri fotosintetizzanti. Pertanto, almeno una parte degli antichi batteri doveva essere dotata di questa capacità. A parte quest'unico fatto, non si potrebbe stabilire praticamente nulla di certo sui primissimi microrganismi. L'intero albero evolutivo dei batteri è rimasto così oscuro, come lo è la base dell'albero relativo alle forme di vita superiori.

Nel ricostruire le prime tappe evolutive, i biologi non si sono limitati però alle testimonianze fossili fornite dalla geologia. La cellula stessa conserva una documentazione del proprio passato nelle sequenze amminoacidiche delle sue proteine e in quelle nucleotidiche degli acidi nucleici DNA e RNA. Questa testimonianza viva è potenzialmente molto più ricca e molto più ampia di qualsiasi testimonianza fossile e risale nel tempo ben oltre i fossili più antichi, fino al periodo in cui esisteva l'antenato comune a tutti gli esseri viventi.

Per poter leggere le testimonianze di ordine biochimico, fu necessario mettere a punto una tecnologia che permettesse di determinare (almeno in parte) la sequenza di un gene o dell'RNA, o del prodotto proteico codificato da un gene. Per le proteine questo è possibile da circa 25 anni, mentre la ricostruzione diretta delle sequenze nucleotidiche del DNA e dell'RNA si è potuta realizzare solo negli ultimi cinque anni. La nuova tecnologia per la definizione delle sequenze negli acidi nucleici dovrebbe permettere ai biologi di scoprire in un tempo relativamente breve molto di più, sulla storia della vita, di quello che si è sempre pensato fosse possibile. Ed è stato proprio applicando tali tecniche al

L'autore propone l'esistenza di tre regni primari, in funzione della scoperta che gli archibatteri sono fondamentalmente diversi da tutti gli altri batteri, designati come eubatteri o batteri veri. Sia gli eubatteri sia gli archibatteri sono cellule procariote, semplici cellule prive di nucleo e con proprietà strutturali molto diverse dalle cellule eucariote, che presentano un nucleo e parecchi altri organelli subcellulari. Da un punto di vista genealogico, però, gli archibatteri e gli eubatteri non sono più affini tra loro di quanto ciascuno di loro lo sia nei riguardi degli eucarioti. Viene avanzata l'ipotesi che gli archibatteri, gli eubatteri e un urcariote - la cellula eucariota originaria - siano tutti derivati da un antenato comune (il progenote), molto più semplice delle più semplici cellule dei nostri giorni (i procarioti). Gli eucarioti, che andarono evolvendosi a partire dall'urcariote, divennero «ospiti» di endosimbionti batterici, che si svilupparono successivamente in mitocondri e cloroplasti.

L'età dei microrganismi, che si estese per circa tre miliardi di anni, domina il tempo dell'evoluzione biologica. Microfossili di cellule procariote sono stati trovati in depositi che risalgono a quel periodo e quelle cellule devono essere state precedute da cellule più semplici. I primissimi fossili degli eucarioti risalgono, invece, a solo un miliardo e trecento milioni di anni fa. Quasi nulla si sa dell'evoluzione e dell'età dei microrganismi. La documentazione fossile macroscopica risale a 600 milioni di anni, all'epoca dei primissimi metazoi o organismi pluricellulari.

problema, ormai secolare, delle relazioni naturali tra i batteri che i miei colleghi e io abbiamo riconosciuto gli archibatteri come una terza forma di vita.

Eucarioti e procarioti

Per poter apprezzare pienamente la condizione particolare degli archibatteri, è utile considerare alcune delle caratteristiche che permettono di definire gli eucarioti e i procarioti. La cellula degli eucarioti è relativamente grossa: misura su un lato pressappoco 10 micrometri. È, inoltre, circondata da una duplice membrana, all'interno della quale si possono distinguere diverse strutture, che a loro volta vengono delimitate da membrane. Il nucleo contiene la maggior parte del materiale genetico della cellula. I mitocondri, di forma bastoncellare, sono la sede della respirazione cellulare, che genera la principale moneta di scambio per l'energia, l'adenosintrifosfato (ATP). Nelle cellule vegetali, il cloroplasto, un altro organello di forma bastoncellare, trasforma l'energia luminosa nell'energia chimica dell'ATP. Altre strutture specializzate, come l'apparato di Golgi (per la secrezione) e le ciglia (per la motilità), sono spesso presenti. Molte cellule eucariote sono dotate di un sistema di membrane, il reticolo endoplasmatico, che offre un'estesa superficie su cui possono svolgersi importanti reazioni, come la sintesi delle proteine.

La cellula dei procarioti è profondamente diversa. Ha una dimensione tipicamente e marcatamente inferiore: di un fattore 10 come misura lineare e, pertanto, di un fattore 1000 come volume. È anch'essa delimitata da una doppia membrana e, inoltre, ha quasi sempre una parete cellulare rigida. D'altra parte, non è presente alcuna delle strutture interne che sono caratteristiche della cellula degli eucarioti; non vi sono né mitocondri né cloroplasti e, naturalmente, non vi è neppure un nucleo avvolto da una membrana. Il genoma, cioè l'intero complesso del materiale genetico, è limitato a 2000-3000 geni, mentre il genoma tipico di una cellula eucariota è maggiore di parecchi ordini di grandezza.

La distinzione tra eucarioti e procarioti è stata posta all'inizio in termini di strutture subcellulari visibili al microscopio. A un simile livello le cellule sembravano o grosse e complesse, quindi eucariote, o piccole e semplici, quindi procariote. Questa distinzione tra i due tipi cellulari è stata spinta alla fine fino al livello biologico più profondo, cioè al livello molecolare, in corrispondenza del quale le cellule procariote e quelle eucariote dimostrano di avere molte caratteristiche in comune. Per esempio, traducono l'informazione genetica in proteine seguendo lo stesso codice genetico. Però, anche laddove i processi molecolari sono gli stessi, i particolari sono diversi nelle due forme: sono o tipicamente eucarioti o tipicamente procarioti. Per esempio, le sequenze amminoacidiche dei vari enzimi tendono a essere o caratteristiche dei procarioti o caratteristiche degli eucarioti. Tutte queste differenze tra gruppi e queste somiglianze all'interno di ogni gruppo hanno fatto credere senza indugi alla maggior parte dei biologi che l'albero della vita avesse due tronchi principali, quello procariota e quello eucariota.

In realtà questa conclusione era stata tratta troppo in fretta: il fascino estetico di una dicotomia aveva prevalso. Non si può dedurre d'acchito che vi sono solo due tipi a livello molecolare semplicemente perché vi sono due tipi di cellule a livello microscopico.

La relazione tra procarioti ed eucarioti sul piano evolutivo è in realtà più complessa di quanto non indichino i fatti sopra esposti. Due organelli degli eucarioti, il mitocondrio e il cloroplasto, possiedono ciascuno il proprio DNA. Inoltre, i pigmenti dei cloroplasti (clorofille e carotenoidi) sono simili a quelli trovati nei cianobatteri. Mitocondri e cloroplasti, inoltre, hanno ambedue le dimensioni dei batteri, mentre il loro apparato di traduzione dell'informazione genetica nelle proteine differisce dall'apparato proprio della cellula degli eucarioti e ha un numero di proprietà in comune con quello dei procarioti.

Queste considerazioni e altre ancora hanno indotto a pensare che mitocondri e

I procarioti e gli eucarioti sono microrganismi fondamentalmente diversi a livello strutturale, come risulta da questi disegni schematici di una tipica cellula procariota *(a sinistra)* e di una eucariota *(a destra)*. Il procariote è di gran lunga la cellula più piccola. Una scarsa struttura subcellulare è visibile anche alla scala del microscopio elettronico: un unico filamento circolare del materiale genetico DNA è libero nel citoplasma. Sia gli archibatteri sia gli eubatteri sono procarioti e hanno in comune le proprietà strutturali di questo gruppo. La cellula eucariota è molto più grossa e presenta un certo numero di strutture subcellulari distinte. Il suo DNA, che forma un complesso con le proteine, è organizzato in cromosomi contenuti all'interno di un nucleo delimitato da una membrana. I mitocondri compiono la respirazione cellulare; nelle cellule vegetali sono presenti i cloroplasti, che effettuano la fotosintesi. L'apparato di Golgi è un organello con funzioni secernenti; il reticolo endoplasmatico è un sistema di membrane, in alcune parti del quale è allineata una parte dei ribosomi della cellula (sui quali l'informazione genetica viene tradotta in proteine). Tutte le cellule che presentano gradi di complessità superiori ai batteri sono eucariote.

L'albero convenzionale della vita aveva due linee di discendenza: quella dei procarioti e quella degli eucarioti, e quest'ultima veniva fatta derivare dalla prima. Si dava come assunto che le prime cellule fossero state batteri anaerobi (procarioti), che ricavavano la loro energia dalla fermentazione. Esse diedero origine a una varietà di linee secondarie di discendenza. E, dopo che l'atmosfera si arricchì di ossigeno, certe cellule anaerobiche che avevano perduto la parete cellulare (micoplasmi) stabilirono un rapporto endosimbiotico con batteri di dimensioni inferiori che avevano ingerito. Un batterio endosimbionte aerobio (cioè che utilizzava l'ossigeno per la respirazione) diede origine evolvendosi a un mitocondrio, un cianobatterio con attività fotosintetica a un cloroplasto e (forse) una spirocheta a un flagello (organello deputato a svolgere funzioni motorie). Così si evolse la cellula eucariota ancestrale, che diede origine ai protozoi, ai funghi, agli animali e ai vegetali. Lo schema è stato proposto da Lynn Margulis della Boston University.

cloroplasti siano derivati da procarioti rimasti intrappolati in una cellula più grossa, con la quale avrebbero stabilito un rapporto endosimbiotico. Si ritiene che il mitocondrio sia stato un batterio con attività respiratoria e il cloroplasto un organismo affine ai cianobatteri, con attività fotosintetica. Questa congettura, che nella sua forma più semplice risale a più di un secolo fa, è stata avvalorata nel caso del cloroplasto dalla dimostrazione che la sequenza nucleotidica di uno dei tipi di RNA presenti in esso, l'RNA ribosomale, ha una specifica correlazione con le sequenze dell'RNA ribosomale dei cianobatteri. Analogamente, l'RNA ribosomale dei mitocondri nelle piante risulta essere di tipo batterico. Sembra in questo modo che nella cellula eucariota siano rappresentate perlomeno due linee di discendenza procariote.

L'urcariote

Logicamente la domanda che ci si pone a questo punto è la seguente: da dove proviene il resto della cellula eucariota? Che cos'era la cellula ospite originaria, l'urcariote? Vi è un consenso generale nel dire che la massa della cellula eucariota (nucleo e strutture citoplamatiche) rappresenta una linea di discendenza separata. La natura esatta della cellula ancestrale non è chiara: la maggior parte dei ricercatori è stata portata a credere che anche la linea principale di discendenza degli eucarioti abbia avuto origine tra i batteri ordinari. L'idea è che qualche batterio anaerobio, che ricavava l'energia dalla fermentazione delle sostanze nutritive (più che dalla loro ossidazione), a un certo punto abbia perduto la rigida parete cellulare. Sono noti organismi di questo tipo: i micoplasmi. Un loro ceppo avrebbe allora sviluppato la capacità di inglobare altri organismi, capacità che si conserva oggi in molti eucarioti. Tra i molti tipi di organismi che un simile micoplasma potrebbe aver ingerito, due sembrano aver stabilito una relazione endosimbiotica stabile con il loro ospite, diventando mitocondri e cloroplasti. In questo modo è nata la cellula eucariota. (L'origine della proprietà che la definisce, cioè il nucleo delimitato da una membrana, è ancora ben lungi dall'essere chiara.)

Quest'interpretazione è soddisfacente per alcuni aspetti, ma non riesce a spiegare le molte differenze tra procarioti ed eucarioti. In particolare, non rende conto dei diversi dettagli dei processi molecolari o delle grosse differenze tra le sequenze amminoacidiche di proteine con analogia funzionale, nei due tipi di cellule. Si dà spesso per scontato che le diversità siano semplicemente una conseguenza dei numerosi piccoli cambiamenti nell'impianto cellulare, che sarebbero stati necessari nel passare dalla semplice condizione dei procarioti alla condizione più complessa degli eucarioti. È discutibile che un numero così elevato di cambiamenti (per esempio cambiamenti nella composizione di quasi tutti gli enzimi) possa ragionevolmente venir spiegato in questo modo.

Soprattutto per questo motivo, alcuni biologi ritengono che la linea di discendenza che ha dato origine alla specie urcariota putativa ha potuto divergere dalla linea dei procarioti in qualche fase precoce, prima della comparsa dello stesso antenato dei batteri. L'urcariote si sarebbe allora potuto evolvere indipendentemente in una forma paragonabile come complessità a quella dei batteri. Una simile ammissione darebbe perlomeno più tempo per emergere alle differenze tra procarioti e urcariote. L'urcariote, allora, rappresenterebbe una linea di discendenza distinta da quella dei procarioti, in accordo con la dicotomia filogenetica fondamentale.

L'RNA ribosomale 16*S* è la molecola le cui sequenze nucleotidiche sono state confrontate in un certo numero di organismi per poter stabilire le loro relazioni filogenetiche. Esso è uno dei componenti del ribosoma, la macchina molecolare che sintetizza le proteine; la designazione 16*S* si riferisce alla velocità, misurata in unità Svedberg (*S*), di sedimentazione in una centrifuga. La molecola di RNA è una lunga catena di subunità, chiamate nucleotidi, ognuna delle quali è caratterizzata da una fra quattro basi: l'adenina (*A*), l'uracile (*U*), la guanina (*G*) e la citosina (*C*). Le prime due basi e le ultime due sono complementari: si possono unire mediante legami idrogeno per formare delle coppie: *A* con *U*, *G* con *C*. L'appaiamento delle basi determina la cosiddetta struttura secondaria della molecola, o il modo in cui comincia a ripiegarsi, formando una cinquantina di strutture brevi a duplice filamento, in cui le basi si appaiano (*zone tratteggiate*). Il disegno mostra la struttura secondaria dell'RNA 16*S* dell'eubatterio *Escherichia coli*, la cui sequenza è stata determinata da H. F. Noller, Jr., della Università della California a Santa Cruz.

Questa era la situazione agli inizi degli anni settanta. La filogenesi degli eucarioti superiori, svoltasi in un arco di circa 500 milioni di anni, è stata compresa piuttosto bene, tranne per quanto riguarda l'importantissima congiunzione dei principali rami degli eucarioti. È emersa un'ipotesi ben definita e largamente accettata, riguardante il modo in cui la cellula eucariota si è evoluta. Le prove in suo favore e le risposte ai restanti interrogativi rimangono, però, nei recessi inesplorati della filogenesi batterica, nell'era dei microrganismi.

L'identificazione delle sequenze genetiche

I batteri rappresentano un mondo di straordinaria varietà, molto più di quanto non riveli il microscopio. Le nicchie ecologiche in cui si trovano superano di gran lunga, come diversità, quelle occupate dalle forme superiori di vita. Per un secolo, i microbiologi hanno cercato invano di capire le relazioni naturali tra batteri e di imporre un certo ordine alla sbalorditiva massa di forme, fisiologie ed ecologie. La varietà tra i batteri è soprattutto una varietà nella semplicità, per cui può dare scarse informazioni sui rapporti filogenetici. Negli organismi superiori, l'occhio, per esempio, si è evoluto un certo numero di volte, ma esso è sufficientemente complicato perché i vari esempi che si sono evoluti in maniera indipendente si possano facilmente distinguere l'uno dall'altro. Questo discorso non può valere, invece, per la forma e la struttura dei batteri: i bastoncini, le sfere e le spirali, che sono le tipiche forme batteriche, si realizzano facilmente e si sono evolute numerose volte. Lo stesso principio vale per la biochimica batterica: anche se alcune caratteristiche dei batteri sono buoni indicatori filogenetici, è impossibile dire in anticipo quali lo sono e quali no.

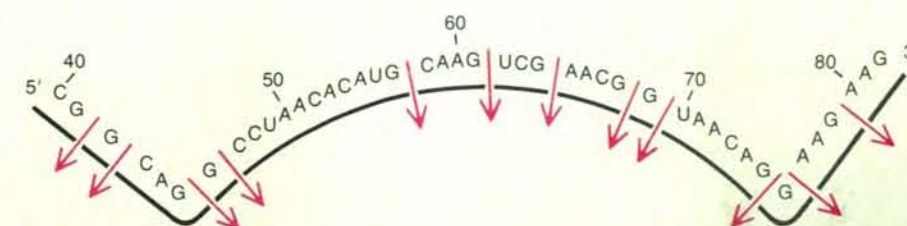
Il modo più semplice in cui la cellula documenta il suo passato è sotto forma di sequenze genetiche. Ogni gene che esiste in una cellula oggi è una copia di un gene che esisteva moltissimi anni fa. Non ne è una copia esatta, perché le mutazioni hanno alterato la sequenza genetica originaria, ma spesso persistono vestigia dello stadio di partenza. Ciò che rende un gene una superba documentazione del passato è la sua semplicità (ha infatti una disposizione lineare) e il fatto che lo «spazio» lasciato a una sequenza genetica sia enorme, per cui nell'intero arco dell'evoluzione si può realizzare solo una piccola percentuale delle possibili sequenze genetiche. Pertanto, se due geni sono simili per un segmento che comprende un numero significativo di nucleotidi, ciò significa solo che essi hanno un antenato in comune; queste molecole, correlate geneticamente, si dicono omologhe.

Una sequenza genetica può dare tre tipi di informazioni sull'evoluzione. Può rivelare dei rapporti genealogici, può misurare il tempo dell'evoluzione e può essere un documento delle caratteristiche ancestrali. Nella misura in cui, in organismi diversi, due geni per la stessa funzione sono correlati, quegli organismi sono affini. E l'entità della differenza tra le sequenze che costituiscono quei due geni dà una valutazione del tempo che è stato necessario perché i due organismi divergessero da un antenato comune. Partendo da un'ampia serie di sequenze correlate, si può costruire un albero filogenetico in cui i punti di ramificazione misurano (approssimativamente) i tempi relativi delle biforcazioni. Infine, i confronti effettuati tra un'ampia serie di sequenze omologhe permettono di ricostruire con una certa precisione varie versioni ancestrali di un gene.

Dato che il rapporto tra un gene e il suo prodotto (o una proteina o uno dei parecchi tipi di RNA) è in genere basato sulla colinearità, la sequenza del prodotto è, di solito, altrettanto utile, per gli studi filogenetici, di quella dello stesso gene. Poiché, fino in epoca recente, solo le sequenze delle proteine potevano essere risolte, proprio partendo da confronti tra queste ultime è stato possibile costruire le prime filogenesi basate sull'evoluzione molecolare. Confronti effettuati nella proteina respiratoria, nota come citrocromo *c*, sono risultati particolarmente preziosi per confermare e ampliare l'albero filogenetico degli organismi superiori. D'altra parte, molecole come quella del citocromo *c* non sono altrettanto efficaci per stabilire relazioni tra i batteri. Esse non sono universalmente distribuite; non hanno funzioni rigorosamente costanti per cui non sono del tutto confrontabili; inoltre, a causa della maggiore antichità delle linee di discendenza batteriche le diversità nelle sequenze possono essere di gran lunga superiori nei batteri rispetto agli eucarioti. Questi fattori rendono incomplete e incerte le filogenesi batteriche dedotte dall'evoluzione delle proteine.

L'RNA ribosomale

Si conoscono altri prodotti genici che possono essere utili come indicatori delle relazioni batteriche. Tutte le entità autoduplicantisi possiedono necessariamente dei sistemi per mantenere e propagare l'informazione genetica e per tradurla in catene di amminoacidi, che costituiscono le proteine. Si deve far risalire l'origine della maggior parte delle grosse molecole impegnate in questi processi ai primissimi stadi dell'evoluzione della cellula; certamente esse comparvero prima che le cellule diventassero tanto complesse da essere chiamate procarioti. Ci sarebbe da supporre, dunque, che esse abbiano i requisiti necessari per essere dei marcatori filogenetici.

Ogni nucleotide di una molecola di RNA si compone di una base, di uno zucchero, il ribosio (*Z*), e di un gruppo fosfato (*P*). Gli enzimi, noti con il nome di ribonucleasi, scindono tale catena in punti specifici. Per esempio, la ribonucleasi T_1 scinde la catena per idrolisi (inserimento di una molecola d'acqua) sul lato 3' del fosfato che segue un qualsiasi nucleotide in cui sia presente guanina. Pertanto scinde una lunga molecola di RNA in un certo numero di frammenti brevi, ciascuno dei quali è formato da uno o più nucleotidi, e termina con un residuo di guanina (*G*).

Viene qui illustrato l'effetto della ribonucleasi T_1 su un breve segmento della molecola di RNA 16*S* di *Escherichia coli*: quando un RNA 16*S* tipico viene scisso in questo modo, la sua sequenza si risolve in frammenti («parole») che variano come lunghezza da un unico nucleotide («lettera») fino a venti nucleotidi. La sequenza di basi di ognuna di queste parole è determinata. Parole di sei lettere o più vengono elencate in un «dizionario» e i dizionari di due organismi possono venir confrontati in termini di coefficiente di associazione S_{AB}. Tale coefficiente è rappresentato da una frazione uguale a due volte il numero di lettere nelle parole (lunghe almeno sei lettere) comuni agli organismi *A* e *B*, diviso per il numero totale di lettere in tutte le parole di questo tipo in *A* e in *B*.

Tra tutte, la cosa più ragionevole sembra far cadere la prima scelta sulle molecole di RNA, che formano dei complessi con le proteine nei ribosomi. Ed è proprio sui ribosomi che l'informazione genetica viene tradotta in proteine. L'RNA ribosomale è facile da isolare in quantità utili perché una cellula batterica tipica contiene da 10 000 a 20 000 ribosomi. Inoltre, sembra che la funzione delle sue molecole si sia mantenuta costante per lunghissimi periodi evolutivi. Questo fatto è importante perché i cambiamenti funzionali in una molecola portano con sé altri cambiamenti, che rendono difficile, se non addirittura impossibile, confrontare una sequenza molecolare con un'altra e dedurre, pertanto, delle relazioni filogenetiche. Un altro vantaggio degli RNA ribosomali è dato dal fatto che perlomeno alcune porzioni delle loro sequenze si modificano abbastanza lentamente per cui la sequenza ancestrale comune non si oblitera del tutto. In altre parole, essi permettono di scoprire i rapporti filogenetici più profondi.

Vi sono tre tipi di molecole di RNA ribosomale. Nei batteri, l'RNA ribosomale «grosso» è l'RNA $23S$ (S è il simbolo dell'unità Svedberg, una misura della velocità di sedimentazione in una ultracentrifuga e pertanto una misura indiretta della dimensione molecolare); esso è lungo all'incirca 2900 nucleotidi. L'unità ribosomale «piccola», designata come RNA $16S$, è lunga, invece, circa 1540 nucleotidi. Un'unità piccolissima ($5S$) ha, infine, solo 120 nucleotidi. Le dimensioni sono simili nelle cellule eucariote: $18S$, 25-$28S$ e $5S$. Si potrebbe pensare che la facilità con cui può venir caratterizzata la piccola unità $5S$ la renda la più adatta per gli studi filogenetici. In realtà, però, essa non è altrettanto precisa come indicatore dei rapporti filogenetici degli RNA ribosomali più grossi, principalmente per ragioni statistiche. (L'RNA ribosomale $5S$ mostra talvolta grosse differenze anomale nella sequenza, da una specie all'altra.) L'RNA ribosomale $16S$ è, invece, la molecola che offre migliori garanzie, mentre la molecola $23S$ ha una dimensione pressoché doppia ed è più che doppiamente difficile da caratterizzare.

I dizionari a RNA

Nel 1969, all'Università dell'Illinois, decisi di studiare i rapporti che esistono tra i batteri, confrontando le sequenze degli RNA ribosomali $16S$ in differenti specie. Non si sapeva ancora determinare (come lo si fa oggi) la sequenza nucleotidica dell'intera molecola. Tuttavia, esisteva già la tecnologia per stabilire le sequenze di corti segmenti. Gli enzimi noti come ribonucleasi producono corti frammenti di RNA, tagliando un filamento di RNA in corrispondenza di punti specifici. Ogni nucleotide dell'RNA consta di uno zucchero chiamato ribosio, di un gruppo fosfato e di una fra quattro basi azotate: l'adenina *(A)*, l'uracile *(U)*, la guanina *(G)* o la citosina *(C)*. L'enzima chiamato ribonucleasi T_1 taglia un filamento di RNA in corrispondenza di un particolare legame su un lato di ogni nucleotide che incorpora una base di guanina. Pertanto, esso scinde un «testo» di RNA in brevi «parole», chiamate oligonucleotidi. Ogni oligonucleotide include e termina con una unica *G*, come in *AACUCG* o in *UCCUAUCG*.

Gli oligonucleotidi ottenuti in questo modo erano sufficientemente corti da poter essere analizzati con le tecniche allora disponibili, in modo da definirne le sequenze. Le parole più brevi hanno uno scarso valore perché ricorrono molte volte in una molecola. Ma quando la lunghezza della parola raggiunge i sei nucleotidi, è improbabile che una particolare sequenza compaia più di una volta in una molecola di RNA $16S$. (Data la base terminale costante *G*, vi sono 3^5, o 243, possibili sequenze a sei lettere di questo tipo e una tipica molecola di RNA $16S$ possiede grosso modo 25 parole del genere.) Quando gli RNA $16S$ di diversi organismi includono la stessa sequenza di sei lettere, quasi sempre ciò significa che esiste tra loro una vera omologia. Limitando l'attenzione alle parole di sei lettere o più, si può compilare un «dizionario», caratteristico di un determinato organismo e facilmente confrontabile con altri dizionari del genere per determinare i rapporti genealogici.

Un modo semplice di analizzare i dati è di esprimerli come coefficiente di associazione S_{AB}, definito come il doppio del numero di nucleotidi nelle parole comuni ad ambedue i dizionari *A* e *B*, diviso per il numero di nucleotidi in tutte le parole presenti nei due dizionari. Il valore di S_{AB} varia da 1 quando i dizionari *A* e *B* sono identici a meno di 0,1 quando non esiste alcun rapporto tra loro. (Il coefficiente è, in generale, superiore a zero anche nel caso di sequenze non correlate e questo a causa di corrispondenze casuali.) Compilando in una tabella i valori di S_{AB} per un certo numero di organismi, si riesce a riconoscere tra questi organismi un certo tipo di affinità o meno. Inoltre, con metodi rigorosamente statistici, è possibile costruire, da una serie di valori di S_{AB} per un gruppo di organismi, un cosiddetto dendrogramma, o albero, che mette in evidenza i rapporti tra i membri del gruppo.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	*Saccharomyces cerevisiae*	—	0,29	0,33	0,05	0,06	0,08	0,09	0,11	0,08	0,11	0,11	0,08	0,08	0,10	0,07	0,08
2	*Lemna minor*	0,29	—	0,36	0,10	0,05	0,06	0,10	0,09	0,11	0,10	0,10	0,13	0,07	0,09	0,07	0,09
3	Cellula *L*	0,33	0,36	—	0,06	0,06	0,07	0,07	0,09	0,06	0,10	0,10	0,09	0,07	0,11	0,06	0,07
4	*Escherichia coli*	0,05	0,10	0,06	—	0,24	0,25	0,28	0,26	0,21	0,11	0,12	0,07	0,12	0,07	0,07	0,09
5	*Chlorobium vibrioforme*	0,06	0,05	0,06	0,24	—	0,22	0,22	0,20	0,19	0,06	0,07	0,06	0,09	0,07	0,05	0,07
6	*Bacillus firmus*	0,08	0,06	0,07	0,25	0,22	—	0,34	0,26	0,20	0,11	0,13	0,06	0,12	0,10	0,07	0,09
7	*Corynebacterium diphteriae*	0,09	0,10	0,07	0,28	0,22	0,34	—	0,23	0,21	0,12	0,12	0,09	0,10	0,10	0,06	0,09
8	*Aphanocapsa*	0,11	0,09	0,09	0,26	0,20	0,26	0,23	—	0,31	0,11	0,11	0,10	0,10	0,13	0,10	0,10
9	Cloroplasto *(Lemna)*	0,08	0,11	0,06	0,21	0,19	0,20	0,21	0,31	—	0,14	0,12	0,10	0,12	0,12	0,06	0,07
10	*Methanobacterium thermoautrophicum*	0,11	0,10	0,10	0,11	0,06	0,11	0,12	0,11	0,14	—	0,51	0,25	0,30	0,34	0,17	0,19
11	*Methanobrevibacter ruminantium*	0,11	0,10	0,10	0,12	0,07	0,13	0,12	0,11	0,12	0,51	—	0,25	0,24	0,31	0,15	0,20
12	*Methanogenium cariaci*	0,08	0,13	0,09	0,07	0,06	0,06	0,09	0,10	0,10	0,25	0,25	—	0,32	0,29	0,13	0,21
13	*Methanosarcina barkeri*	0,08	0,07	0,07	0,12	0,09	0,12	0,10	0,10	0,12	0,30	0,24	0,32	—	0,28	0,16	0,23
14	*Halobacterium halobium*	0,10	0,09	0,11	0,07	0,07	0,10	0,10	0,13	0,12	0,34	0,31	0,29	0,28	—	0,19	9,23
15	*Sulfolobus acidocaldarius*	0,07	0,07	0,06	0,07	0,05	0,07	0,06	0,10	0,06	0,17	0,15	0,13	0,16	0,19	—	0,13
16	*Thermoplasma acidophilum*	0,08	0,09	0,07	0,09	0,07	0,09	0,09	0,10	0,07	0,19	0,20	0,21	0,23	0,23	0,13	—

Questa tabella dei coefficienti di associazione mostra l'affinità tra due organismi qualsiasi, che è maggiore quanto più alto è il valore della frazione S_{AB}. L'insieme è molto significativo. Gli eucarioti *(1-3)*, gli eubatteri *(4-9)* e gli archibatteri *(10-16)* formano ciascuno un gruppo distinto *(in colore)*. Gli archibatteri non sono più vicini agli eubatteri di quanto lo siano agli eucarioti. *Saccharomyces* è un lievito; *Lemna* una pianta acquatica; le cellule *L* sono una linea cellulare di topo. Il cloroplasto deriva da un cianobatterio endosimbionte ed è di origine eubatterica.

Fino a oggi sono stati caratterizzati gli RNA ribosomali di quasi duecento specie di batteri e di eucarioti. La maggior parte dei batteri forma un gruppo coerente, ma molto grande (il che vuol dire antico). Si tratta di eubatteri, o batteri veri e, come c'è da supporre, ben distinti dagli eucarioti. I rapporti tra i vari generi (rappresentati dalle ramificazioni dell'albero) sono determinati mediante un'analisi degli RNA ribosomali e risultano in disaccordo con molti dei pregiudizi ormai radicati, riguardanti i rapporti tra i batteri. Ciò che importa a questo punto è che gli eubatteri siano suddivisi in un certo numero di importanti ramificazioni e che parecchie di queste includano dei batteri con attività fotosintetica. Questo risultato suggerisce l'idea che tutti gli eubatteri derivino da un antenato comune, dotato di attività fotosintetica.

I metanobatteri, batteri anaerobi che producono metano (CH_4) partendo dall'idrogeno e dall'anidride carbonica, costituiscono il gruppo più numeroso di archibatteri che sia stato finora identificato. In queste microfotografie al microscopio elettronico a scansione, scattate da Alexander J. B. Zehnder del Politecnico Federale di Zurigo, sono visibili quattro generi che hanno dimensioni e morfologia profondamente diverse. Si tratta di *Methanosarcina (in alto a sinistra)*, di *Methanobrevibacter (in alto a destra)*, di *Methanospirillum (in basso a sinistra)* e di *Methanobacterium (in basso a destra)*. Le loro cellule risultano ingrandite rispettivamente di 2500, 5000, 1000 e ancora 5000 diametri. I metanobatteri si trovano esclusivamente in ambienti privi di ossigeno.

La scoperta degli archibatteri

Proseguendo nell'esame dei batteri, ebbi una sorpresa. In collaborazione con Ralph S. Wolfe, incominciai a osservare l'RNA ribosomale dei metanobatteri. Questi insoliti organismi vivono solo in ambienti privi di ossigeno e generano metano (CH_4) per riduzione dell'anidride carbonica (CO_2). Scoprimmo che i batteri metanogeni non rientrano nel gruppo filogenetico definito dagli altri batteri e, in effetti, essi sembrano rappresentare un ramo evolutivo che ha preceduto di gran lunga l'antenato comune di tutti i batteri veri. Non solo essi erano separati, ma il gruppo che costituivano sembrava essere altrettanto antico, filogeneticamente, del gruppo definito dagli eubatteri.

Non sembrano esservi dubbi che i batteri metanogeni e i loro affini siano batteri. Hanno le dimensioni dei batteri, non hanno membrana nucleare, hanno un basso contenuto di DNA, e così via. Ci sarebbe da aspettarsi, allora, una loro maggiore affinità con gli altri batteri piuttosto che con gli eucarioti. La nostra analisi ha mostrato il contrario o perlomeno che i metanobatteri sono affini in pari modo agli eucarioti e agli eubatteri.

Com'è possibile? Si supponeva che vi fossero solo due linee di discendenza primarie, quella degli eucarioti e quella dei procarioti. Ed ecco che si aveva a che fare con un nuovo gruppo di organismi: i metanobatteri e loro affini, che complessivamente erano stati chiamati archibatteri. Erano ovviamente simili agli altri batteri per le caratteristiche superficiali e, per questo motivo, si era ammesso che facessero parte della linea di discendenza dei procarioti. Non sono però delle spettacolari differenze nelle caratteristiche morfologiche a permettere di distinguere filogeneticamente la cellula dei procarioti da quella degli eucarioti; sono, invece, delle differenze più sottili e più antiche nelle sequenze molecolari e nei particolari delle funzioni a livello molecolare. Pertanto non vi è alcuna ragione per cui due linee di discendenza procariote non possano essere altrettanto distinte tra loro quanto l'una o l'altra lo è dalla linea di discendenza eucariota.

Quest'idea era troppo insolita per poter essere accolta facilmente e, all'inizio, alcuni biologi rifiutarono la possibilità di una «terza forma di vita». Come poteva un batterio non essere qualcosa che sembrava un batterio ma che, di fatto, non era correlato con i batteri? Col passare del tempo, la semplicità della nostra argomentazione e l'ammassarsi di prove

In questa microfotografia al microscopio elettronico a scansione, scattata da Frank Mayer dell'Università di Gottinga, un'insolita parete cellulare di *Methanogenium marisnigri*, un metanobatterio che è stato ricuperato sul fondale del Mar Nero, appare ingrandita di 70 000 diametri. La disposizione a mosaico delle subunità di natura proteica è tipica di parecchi archibatteri. Essa è diversa dalla parete cellulare tipica degli eubatteri formata da subunità costituite da peptidoglicano, le quali non entrano nella composizione della parete cellulare degli archibatteri.

riuscirono ad avere il sopravvento. Anche se ancora oggi alcuni biologi discutono sulla nostra interpretazione, l'idea che gli archibatteri rappresentino un raggruppamento separato al livello più alto viene gradatamente accettata con un consenso sempre più generale.

La supposta grande antichità degli archibatteri rimane un pregiudizio non dimostrato, ma plausibile. Il fenotipo del metanobatterio sembra coprire, nella filogenesi, una distanza altrettanto grande, o ancora più grande, di quella coperta da qualsiasi altro fenotipo batterico a esso paragonabile. Ciò implica che i metanobatteri sono altrettanto antichi, o anche più antichi, di qualsiasi altro gruppo batterico. Inoltre, il loro metabolismo (riduzione di anidride carbonica in metano) è teoricamente adatto al tipo di atmosfera che si ritiene sia esistita sulla Terra primitiva: un'atmosfera ricca di anidride carbonica, con un poco di idrogeno, ma praticamente senza ossigeno. Il nome di archibatteri sottintende che questi organismi fossero dominanti nella biosfera primeva. Quando le condizioni si modificarono, la loro esigenza di vivere in un ambiente anaerobio li confinò in una gamma limitata di nicchie relativamente inaccessibili.

Le misurazioni che rivelarono l'esistenza degli archibatteri (differenze nelle sequenze degli RNA) erano genetiche ed esclusivamente quantitative. Non rivelarono nulla riguardo alla qualità delle differenze - differenze fenotipiche - tra gli archibatteri e i batteri veri. Se la nostra interpretazione degli archibatteri come regno separato da quello dei batteri veri è corretta, allora, in seguito a un esame particolareggiato, gli archibatteri dovrebbero risultare altrettanto diversi nel loro fenotipo molecolare dai batteri veri di quanto l'uno e l'altro gruppo lo sono dalle cellule degli eucarioti.

Forme di archibatteri

Gli archibatteri sono, effettivamente, organismi insoliti. È noto oggi che comprendono tre tipi molto differenti di batteri: i metanobatteri, gli alofili estremi e i termoacidofili.

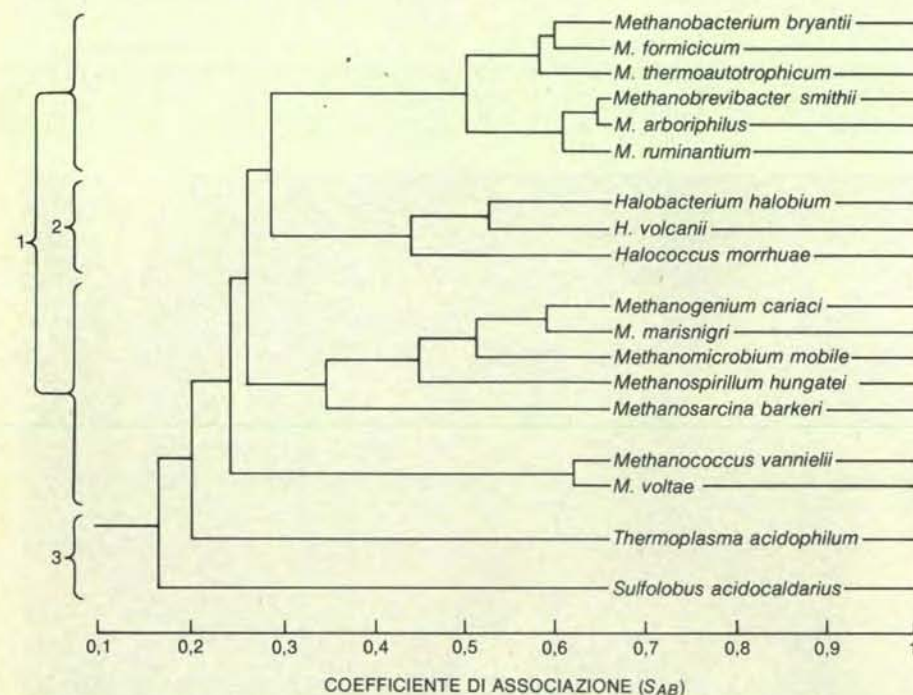
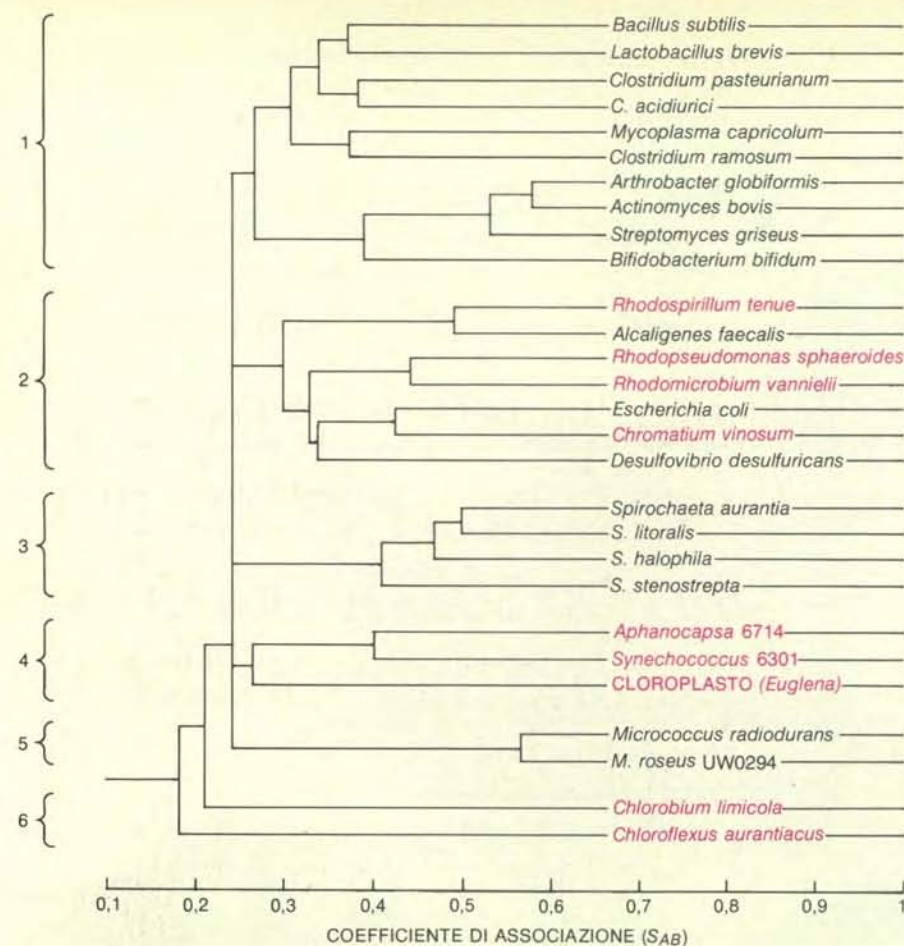
La forma dominante (nel senso che costituisce un raggruppamento filogenetico molto esteso) è quella dei metanobatteri. Da parecchio tempo erano noti batteri che emettevano metano. Nel 1776, Alessandro Volta scoprì che negli stagni acidi si formava dell'«aria combustibile», e così pure nei corsi d'acqua e nei laghi i cui sedimenti sono ricchi di materiale vegetale in decomposizione. Il fatto che un microrganismo fosse responsabile della produzione del «gas delle paludi» fu acquisito però solo molto tempo dopo. I metanobatteri hanno un'ampia distribuzione in natura, ma non si incontrano molto spesso perché vengono uccisi dall'ossigeno e non possono sopravvivere all'aperto.

Nei tempi antichi essi potevano esistere pressoché dovunque. Oggi vivono solo dove l'ossigeno è escluso e dove, al suo posto, si trovano, invece, idrogeno e anidride carbonica. Questo significa, in generale, vivere in stretta associazione con altri batteri, quale *Clostridium*, che metabolizzano la sostanza organica in decomposizione e producono idrogeno come scoria. I metanobatteri si trovano nelle acque stagnanti e negli impianti di trattamento degli scarichi urbani (in quantità che li hanno resi sfruttabili commercialmente per produrre metano). Si trovano anche nel rumine dei bovini e di altri ruminanti e nell'intestino degli animali in genere. Possono essere isolati dai fondali marini e dalle acque delle sorgenti calde. Malgrado la loro non tolleranza per l'ossigeno, hanno una evidente distribuzione globale.

Gli alofili estremi sono batteri che necessitano di forti concentrazioni di sale per poter sopravvivere; alcuni crescono senza difficoltà in acque salmastre sature. Possono conferire una colorazione rossa agli stagni di evaporazione del sale e possono, invece, far perdere il colore al pesce salato, deprezzandolo. Questi microrganismi crescono in habitat salati lungo le coste marine e nelle acque interne come il Grande Lago Salato e il Mar Morto. Pur essendo da lungo tempo oggetto di studio da parte dei microbiologi, solo di recente si è avuto per essi un interesse particolare. E questo per due ragioni: attraverso la loro membrana cellulare mantengono dei notevoli gradienti nella concentrazione di certi ioni e sfruttano questi gradienti per far passare varie sostanze dentro e fuori la cellula. Inoltre gli alofili estremi hanno un meccanismo fotosintetico relativamente semplice, basato non sulla clorofilla, ma su un pigmento, la rodopsina batterica, che è legato alla membrana ed è notevolmente simile a uno dei pigmenti visivi.

I batteri termoacidofili

Il terzo tipo di archibatteri è rappresentato dai termoacidofili, anch'essi notevoli per l'habitat. *Sulfolobus*, uno dei due generi in cui si suddividono, si trova in sorgenti solforose calde. Le sue varie specie crescono generalmente a temperature vicine agli 80 °C, mentre per alcune varietà è stata osservata una crescita addirittura a temperature superiori ai 90 °C. Inoltre, le sorgenti nelle quali prospera *Sulfolobus* sono estremamente acide: sono comuni in esse valori di *p*H inferiori a 2 (la neutralità è a *p*H 7). *Thermoplasma*, il secondo genere dei termoacidofili, è stato trovato finora solo nei mucchi di residui del carbone, che bruciano senza fiamma. Si tratta di un micoplasma: non ha parete cellulare, ma semplicemente una membrana cellulare limitante.

Anche se gli archibatteri termoacidofili possono crescere solo in ambiente acido, il mezzo interno della loro cellula ha un *p*H moderato, vicino alla neutralità. Ciò esige che un gradiente misurabile di *p*H si conservi attraverso la membrana cellulare. Come nel caso degli alofili estremi, esso può svolgere un ruolo nel

Il dendrogramma, o albero, degli archibatteri è stato ottenuto partendo dai valori di S_{AB} e mostra le relazioni filogenetiche tra i membri di questo regno primario. La maggior parte sono metanobatteri *(1)*, che sono anaerobi (cioè organismi che sopravvivono solo in assenza di ossigeno) e generano metano per riduzione dell'anidride carbonica o di certe altre fonti di carbonio molto semplici. Gli alofili estremi *(2)* sono, invece, batteri aerobi (cioè che si servono dell'ossigeno per la respirazione), che vivono solo in ambienti con una concentrazione salina elevata. I termoacidofili *(3)* sono anch'essi batteri aerobi che vivono solo in ambienti estremamente acidi e a temperature molto elevate. I metanobatteri sembrano essere un gruppo antico all'interno del quale sarebbero comparsi gli alofili; i termoacidofili potrebbero, invece, essere comparsi separatamente.

Il dendrogramma degli eubatteri mostra sei principali sottogruppi, tre dei quali includono dei batteri dotati di attività fotosintetica *(in colore)*; altri gruppi devono ancora essere definiti. I batteri Gram-positivi *(1)* possiedono una spessa parete cellulare con una composizione unica che assorbe e trattiene la colorazione di Gram. I batteri purpurei con attività fotosintetica sono raggruppati *(2)* con un numero di specie strettamente affini, ma privi di tale attività, avendo presumibilmente perso questa prerogativa ancestrale. Le spirochete *(3)* sono lunghi batteri spiralati. I cianobatteri *(4)* sono dotati di attività fotosintetica e producono ossigeno; i cloroplasti sono derivati da loro. Alcuni batteri di forma sferica con una parete cellulare atipica *(5)* hanno una singolare resistenza alle radiazioni. I batteri verdi con attività fotosintetica *(6)* sono anaerobi.

pompare altre molecole verso l'interno e verso l'esterno della cellula. È interessante che, quando la temperatura si abbassa e, di conseguenza, *Sulfolobus* cessa di metabolizzare, il *p*H interno non può più essere mantenuto vicino a valori neutri e la cellula muore.

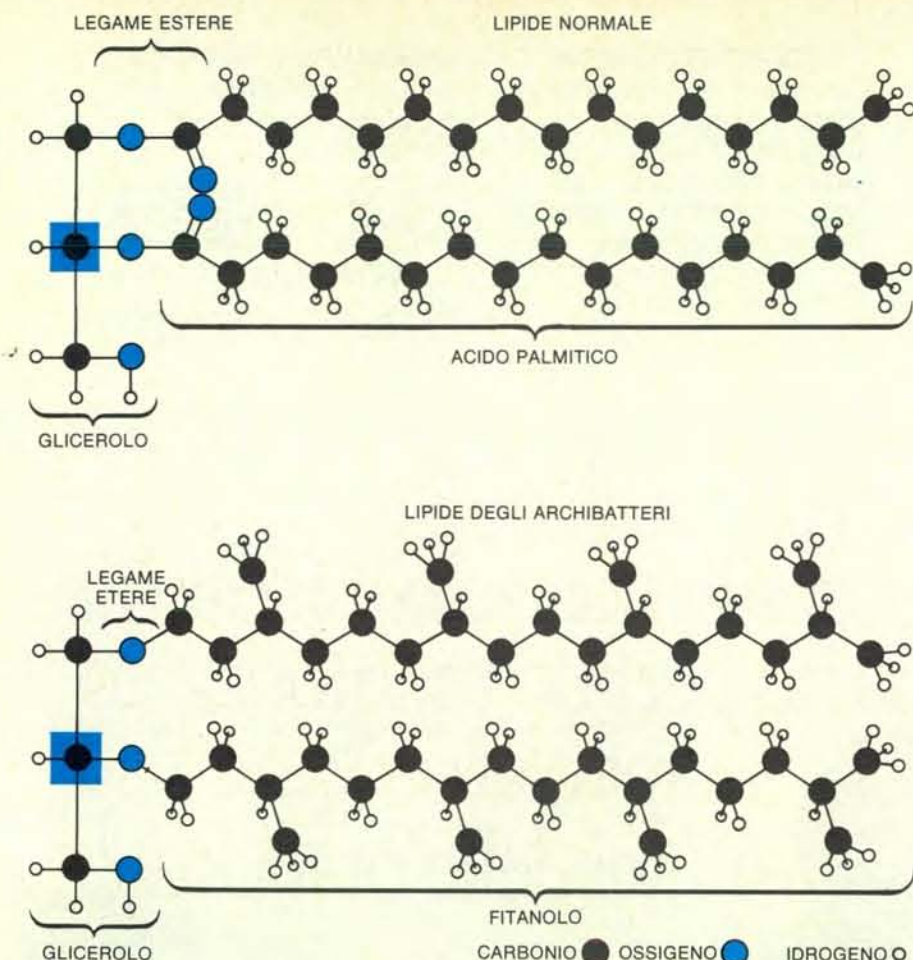
È stato riconosciuto, di recente, che i vari organismi oggi classificati come archibatteri sono individualmente qualcosa di peculiare. In ogni caso la peculiarità è stata vista esattamente così: un adattamento a qualche nicchia caratteristica o una scappatoia biochimica. La misurazione filogenetica dell'RNA ribosomale, invece, ha mostrato che perlomeno alcune delle peculiarità potevano essere caratteristiche generali di un nuovo gruppo di organismi. Con queste cognizioni, i ricercatori di molti paesi hanno cominciato a trovare le proprietà generali che uniscono gli archibatteri l'uno all'altro e a vedere in che modo queste proprietà o distinguono gli archibatteri dalle altre due forme principali oppure li correlano in maniera specifica all'una o all'altra di queste forme.

Si è potuto riconoscere un carattere generale a tutti i batteri, che è quello di possedere una parete cellulare che incorpora l'acido muramico, un derivato degli zuccheri, e unità di base di un polimero complesso noto come peptidoglicano. Ma si è anche saputo che un alofilo estremo e il termoacidofilo *Sulfolobus* rappresentavano delle eccezioni a questa regola, avendo presumibilmente una struttura della parete del tutto particolare. Otto Kandler dell'Università di Monaco, in collaborazione con Wolfe, ha compiuto uno studio sistematico sulla struttura della parete cellulare in altri archibatteri noti ed è risultato che, in tutti, essa era atipica. Gli archibatteri hanno cioè tutta una varietà di tipi di parete, ma nessuna è fatta di peptidoglicano con unità costitutive di acido muramico.

I lipidi e gli RNA

Un'altra nozione che venne acquisita fu che la membrana cellulare degli acidofili estremi e dei termoacidofili si compone di lipidi insoliti. I lipidi sia degli eucarioti, sia degli eubatteri consistono per la maggior parte di due acidi grassi a catena lineare, legati a una estremità con una molecola di glicerolo attraverso un legame estere (–CO–O–). I lipidi degli alofili estremi e dei termoacidofili sono anch'essi costituiti di glicerolo legato a due lunghe catene di idrocarburi, ma il legame tra glicerolo e le catene è del tipo etere (–O–) e non estere. Inoltre, le catene di idrocarburi non sono lineari, ma ramificate: ogni quattro atomi di carbonio la catena presenta un gruppo metilico (CH_3). In altre parole il lipide di base negli archibatteri è un dietere composto da glicerolo e da due molecole di un alcool, il fitanolo. Quando si potè esaminare un certo numero di metanobatteri per conoscere la loro composizione lipidica, le nostre attese furono confermate: i lipidi di questi organismi risultarono essere eteri del glicerolo, con catene ramificate.

Nel corso degli studi sull'RNA ribosomale, emerse un'altra inattesa proprietà degli archibatteri, che avrebbe fornito il primo chiarimento sul significato delle differenze tra archibatteri e batteri veri. Punto focale del processo di traduzione è la molecola di RNA di trasporto, che riconosce nell'RNA messaggero un codone di tre basi, che specifica un particolare amminoacido, e cede quell'amminoacido in modo che sia incorporato nella nascente catena proteica. Un certo numero dei nucleotidi che compongono la molecola dell'RNA di trasporto vengono modificati, cioè la loro struttura viene alterata chimicamente, dopo che sono stati incorporati nella molecola; il più delle volte, un gruppo metilico si aggiunge al nucleotide in corrispondenza di qualche punto sia sulla base sia sullo zucchero. Si ritiene oggi in biologia che in quasi tutti gli organismi e in quasi tutte le molecole di RNA di trasporto, una particolare modificazione sia caratteristica di una certa posizione: in quella posizione la base uracile è stata metilata per formare la timina, che normalmente è presente solo nel DNA e non nell'RNA. Ne deriva che tutti gli RNA di trasporto di tutti gli archibatteri sono privi di questa unità di timina; mentre, al contrario, l'uracile è stato modificato in uno dei due altri modi possibili: a dare una pseudouridina o un nucleotide finora non identificato.

Se si confrontano sia l'RNA ribosomale sia l'RNA di trasporto negli eucarioti, negli eubatteri e negli archibatteri, si trova uno schema generale, di cui la sostituzione della timina negli RNA di trasporto degli archibatteri è solo un esempio. Le stesse regioni negli RNA tendono a essere modificate in tutte e tre le linee di discendenza primarie, ma la natura della modificazione varia, invece, da un regno all'altro. Le differenze sono di due tipi. O la modificazione di una data base è diversa in ciascuno dei regni, oppure una determinata base viene modificata in un regno, mentre in un altro regno la modificazione avviene su una base adiacente. Queste modalità di variazione suggeriscono che, in ogni principale linea di discendenza, le modificazioni si sono evolute separatamente.

I lipidi delle membrane degli archibatteri sono diversi dai lipidi che si trovano in altri organismi. I lipidi degli eubatteri e degli eucarioti sono esteri del glicerolo con acidi grassi a catena lineare. In altre parole, quindi, sono composti da un alcool a tre atomi di carbonio, il glicerolo appunto, che si lega a catene di acidi grassi come l'acido palmitico, mediante un legame estere (–CO–O–). I lipidi degli archibatteri, invece, sono dieteri in cui una unità di glicerolo si unisce mediante legame etereo (–O–) ai fitanoli, catene ramificate in cui gli atomi di carbonio a intervalli regolari portano un gruppo metilico (CH_3). Inoltre, il glicerolo ha due isomeri ottici, che si distinguono per la configurazione della molecola attorno all'atomo di carbonio centrale *(riquadro in colore)*; gli isomeri ottici fanno ruotare il piano della luce polarizzata in direzione opposta. La configurazione del glicerolo attorno all'atomo di carbonio centrale, che si trova nei lipidi degli archibatteri, è l'immagine speculare della configurazione riscontrabile sia nei lipidi degli eubatteri sia in quelli degli eucarioti.

Parecchie altre differenze molecolari tra gli archibatteri e gli altri gruppi sono note (per esempio, nella struttura delle subunità dell'enzima RNA-polimerasi), ma la lista non è lunga. La ragione non è che non esistano altre differenze, ma piuttosto che il mondo degli archibatteri rimane praticamente inesplorato. Lo studio della genetica degli archibatteri è a uno stadio ancora primitivo: sono stati isolati finora solo pochi mutanti per l'indagine genetica. Nulla poi si sa riguardo al controllo dell'espressione genica negli stessi. Non si conosce la biologia molecolare di base per questo gruppo eppure, nella misura in cui sono stati caratterizzati, si è trovato che gli archibatteri differiscono in maniera significativa da ambedue gli altri gruppi principali.

Una nuova prospettiva

La scoperta di un nuovo regno primario di organismi è un risultato importante in sé (paragonabile all'andare in cortile e vedere un organismo che non è né un animale né una pianta), ma la reale importanza della scoperta sta in ciò che può rivelare riguardo alla storia iniziale della vita. Quando si conoscevano solo due linee primarie di discendenza, non si riuscivano a interpretare facilmente le differenze tra le due. L'aver riconosciuto tre linee di discendenza equidistanti l'una dall'altra ha dato una prospettiva molto migliore per poter giudicare quali proprietà sono ancestrali e quali, invece, si sono evolute di recente. Con la scoperta degli archibatteri, due problemi centrali dell'evoluzione si sono pertanto potuti affrontare: la natura dell'antenato comune a tutte le forme di vita e l'evoluzione della cellula eucariota.

A quale stadio nell'evoluzione della cellula ebbe luogo la divisione fondamentale nei regni primari? Qual è la natura dell'antenato universale? Si è ammesso che quest'antenato universale fosse un procariote, la più semplice delle forme viventi attuali. Molto tempo addietro, invece, devono esserci state forme cellulari ancora più semplici. Anche se nulla si sa di esse, si possono avanzare prudenti supposizioni su certune delle proprietà generali.

Si consideri la seguente argomentazione. Il meccanismo di traduzione è complesso comprendendo un certo numero di grossi componenti molecolari, nell'ordine del centinaio. Esso ha anche un funzionamento estremamente accurato. Nella sintesi di proteine delle dimensioni oggi comuni (catene costituite da 100 a 500 amminoacidi), il fatto di ottenere un prodotto perfetto il 90 per cento delle volte o più, corrisponde a un tasso di errore non superiore ad alcune unità su diecimila. Inoltre, questa precisione deve essere raggiunta tramite un meccanismo di dimensioni molecolari. È chiaramente impossibile che esso si sia evoluto in un'unica fase. La sua versione primitiva deve essere stata molto più semplice, di dimensioni più limitate e meno precise.

Una imprecisione nella traduzione avrebbe significato la sintesi di proteine più piccole, e pertanto con un'azione meno specifica, di quelle attuali. (Altrimenti la probabilità di errore nel sintetizzare un filamento proteico sarebbe stata troppo grande.) Tra le proteine di minori dimensioni e meno specifiche vi sarebbero stati gli enzimi necessari per elaborare l'informazione genetica. Se questi enzimi fossero stati meno precisi delle versioni attuali, il tasso di mutazione a carico della cellula sarebbe stato necessariamente più elevato e, di conseguenza, le dimensioni del suo genoma più piccole. Il processo di traduzione è il legame tra genotipo e fenotipo, tra informazione genetica e la sua espressione. A mano a mano che esso si è evoluto per diventare più preciso, la cellula è passata necessariamente attraverso una serie corrispondente di perfezionamenti evolutivi. Si è evoluta - a partire da una entità che ha proprietà semplici, funzioni generiche e imprecise e un corredo piuttosto piccolo di geni - una entità che funziona con numerosi enzimi, dotati di una specificità elevata, e un apparato genetico preciso e complesso. Per mettere meglio in evidenza i meccanismi genetici e di traduzione delle cellule primitive e più semplici, chiamerò queste ultime «progenoti».

I progenoti come antenati

La scoperta degli archibatteri dà la prospettiva necessaria per poter affrontare la questione se l'antenato universale fosse un procariote o un progenote. Benché essa sia ben lungi dall'essere stata risolta, le prime indicazioni confermano che l'antenato universale era in effetti un progenote. Prima di tutto si consideri che i batteri veri e gli archibatteri esistono probabilmente da almeno 3 miliardi e mezzo di anni. Pertanto il tempo necessario per l'evoluzione dei primi veri batteri o archibatteri deve essere stato inferiore al miliardo di anni, e forse molto meno.

	ARCHIBATTERI	EUBATTERI	EUCARIOTI
DIMENSIONI DELLA CELLULA (DIMENSIONE LINEARE)	CIRCA UN MICROMETRO	CIRCA UN MICROMETRO	CIRCA 10 MICROMETRI
ORGANELLI CELLULARI	ASSENTI	ASSENTI	PRESENTI
MEMBRANA NUCLEARE	ASSENTE	ASSENTE	PRESENTE
PARETE CELLULARE	VARIETÀ DI TIPI; NESSUNO INCORPORA L'ACIDO MURAMICO	VARIETÀ ALL'INTERNO DI UN UNICO TIPO; TUTTI INCORPORANO ACIDO MURAMICO	ASSENTE NELLE CELLULE ANIMALI; VARIETÀ DI TIPI NEGLI ALTRI PHYLA
LIPIDI DELLE MEMBRANE	CATENE ALIFATICHE RAMIFICATE CON LEGAMI ETERE	CATENE ALIFATICHE LINEARI CON LEGAMI ESTERE	CATENE ALIFATICHE LINEARI CON LEGAMI ESTERE
RNA DI TRASPORTO:			
TIMINA NEL BRACCIO «COMUNE»	ASSENTE	PRESENTE NELLA MAGGIOR PARTE DEGLI RNA DI TRASPORTO DELLA MAGGIOR PARTE DELLE SPECIE	PRESENTE NELLA MAGGIOR PARTE DEGLI RNA DI TRASPORTO DI TUTTE LE SPECIE
DIIDROURACILE	ASSENTE IN TUTTI I GENERI TRANNE UNO	PRESENTE NELLA MAGGIOR PARTE DEGLI RNA DI TRASPORTO DI TUTTE LE SPECIE	PRESENTE NELLA MAGGIOR PARTE DEGLI RNA DI TRASPORTO DI TUTTE LE SPECIE
AMMINOACIDO VEICOLATO DALL'RNA DI TRASPORTO INIZIATORE	METIONINA	FORMILMETIONINA	METIONINA
RIBOSOMI:			
DIMENSIONI DELLE SUBUNITÀ	30*S*, 50*S*	30*S*, 50*S*	40*S*, 60*S*
LUNGHEZZA APPROSSIMATIVA DELL'RNA 16*S* (18*S*)	1500 NUCLEOTIDI	1500 NUCLEOTIDI	1800 NUCLEOTIDI
LUNGHEZZA APPROSSIMATIVA DELL'RNA 23*S* (25-28 *S*)	2900 NUCLEOTIDI	2900 NUCLEOTIDI	3500 NUCLEOTIDI O PIÙ
FATTORE DI ALLUNGAMENTO NELLA TRADUZIONE	REAGISCE CON LA TOSSINA DIFTERICA	NON REAGISCE CON LA TOSSINA DIFTERICA	REAGISCE CON LA TOSSINA DIFTERICA
SENSIBILITÀ AL CLORAMFENICOLO	NESSUNA	BUONA	NESSUNA
SENSIBILITÀ ALLA ANISOMICINA	BUONA	NESSUNA	BUONA
SENSIBILITÀ ALLA CANAMICINA	NESSUNA	BUONA	NESSUNA
SITO DI LEGAME PER L'RNA MESSAGGERO *AUCACCUCC* IN CORRISPONDENZA DELLA ESTREMITÀ 3' DELL'RNA 16*S* (18*S*)	PRESENTE	PRESENTE	ASSENTE

I caratteri molecolari di archibatteri, eubatteri ed eucarioti distinguono nettamente questi tre regni primari. In alcuni casi gli archibatteri sono unici; in altri sono simili o agli eubatteri o agli eucarioti, come mostra la distribuzione del colore. La timina e il diidrouracile sono basi modificate che sostituiscono l'uracile negli RNA di trasporto. Il fattore di allungamento è un componente del meccanismo di traduzione.

Inoltre, i tipi di cambiamenti che sono insorti all'interno di ogni regno batterico, nell'intervallo di tempo successivo di tre miliardi di anni o più, sono di minore importanza in confronto alle differenze che separano gli archibatteri dai veri batteri: per esempio, differenze nei lipidi, nelle sequenze e nel tipo di modificazioni degli RNA di trasporto e degli RNA ribosomali, infine nella struttura delle subunità enzimatiche.

Sembrerebbe che la natura dell'evoluzione circa quattro miliardi di anni fa fosse molto diversa da quella che si è avuta in seguito. Ciò implica che anche gli organismi soggetti a evoluzione erano molto diversi. La possibilità che l'antenato universale stesse elaborando la parete cellulare è suggerita dal fatto che la cellula archibatterica ha pareti che sono tanto diverse dalle pareti dei batteri veri quanto da quelle delle cellule eucariote. Può darsi anche che tale antenato stesse ancora sviluppando o perfezionando le vie biochimiche; i lipidi sono sintetizzati in maniera diversa nei due regni batterici e molti coenzimi sono diversi. Se si dovessero trovare degli archibatteri che differiscono dai batteri veri nei meccanismi di controllo dell'azione genica (una possibilità che deve ancora essere analizzata), se ne dedurrebbe che il loro antenato comune può aver avuto solo meccanismi rudimentali di controllo genetico.

Il problema chiave è di vedere se l'antenato universale stava ancora sviluppando il legame genotipo-fenotipo quando diede origine alle linee discendenti. Due osservazioni fanno pensare di sì. La RNA-polimerasi è l'enzima che trascrive il gene nell'RNA messaggero (il quale viene successivamente tradotto in proteina). La struttura in subunità della RNA-polimerasi è costante nei batteri veri, mentre la polimerasi degli archibatteri ha una struttura differente. Significherebbe forse questo che la funzione della RNA-polimerasi stava ancora perfezionandosi quando le linee di discendenza batteriche si separarono?

La seconda osservazione riguarda i nucleotidi modificati negli RNA di trasporto e in quelli ribosomali. Come ho già ricordato, gli schemi di modificazione sono praticamente invarianti all'interno di un qualsiasi regno primario, mentre tendono a essere differenti tra i diversi regni. Non si comprende la funzione dei nucleotidi modificati nei due tipi di RNA sopra menzionati, ma è ragionevole ammettere che la maggior parte serva a rendere la traduzione più precisa, a sintonizzarla meglio. Se fosse proprio così, ne risulterebbe che in ogni regno primario si sono evolute indipendentemente molte modificazioni. Questa indipendenza implica a sua volta che l'antenato universale non possedeva le molecole altamente specializzate di oggi per l'RNA di trasporto e l'RNA ribosomale, ma se la cavava con un meccanismo di traduzione più rudimentale. A questo stadio, si può dire solo che i fatti concordano con l'idea, e anzi la suggeriscono, che l'antenato universale abbia avuto meccanismi di trascrizione e di traduzione rudimentali e che sia stato pertanto un progenote.

L'origine dell'urcariote

Gli archibatteri, gli eubatteri e gli eucarioti sembrano essere genealogicamente equidistanti tra loro se si considerano i loro RNA ribosomali; non è stata individuata alcuna relazione specifica tra due qualsiasi di loro. Cionondimeno, se si considera la sequenza amminoacidica di una proteina, la proteina *A* ribosomale, gli archibatteri sembrano chiaramente affini agli eucarioti. Pertanto può darsi che sia gli archibatteri sia i batteri veri abbiano partecipato alla formazione della cellula eucariota. Forse bisognerebbe guardare proprio agli archibatteri per l'origine dello stadio non ancora spiegato della cellula eucariota: l'urcariote che svolse il ruolo di ospite nei riguardi degli antenati endosimbionti dei mitocondri e dei cloroplasti. (Piuttosto che cercare l'ipotetico ospite, però, ci si dovrebbe a questo punto chiedere se una simile entità sia mai esistita. Non è questo il tempo di modellare nuove scoperte su vecchi pregiudizi.)

Come ho già detto, le differenze che esistono tra la cellula eucariota e gli altri principali tipi di cellule a livello molecolare sono molto più estese e profonde di qualunque differenza che sia visibile al microscopio. Il nucleo degli eucarioti sembra contenere perlomeno tre tipi di geni: quelli di origine eubatterica (si presume che il nucleo se ne sia appropriato dai genomi degli organelli presenti nell'eucariote), quelli di origine archibatterica (per esempio il gene che controlla la sintesi della proteina *A* ribosomale) e quelli di origine non identificata e comunque diversa dalle precedenti (ne sono esempi quelli che codificano per l'RNA ribosomale citoplasmatico). In quale misura il nucleo degli eucarioti è una chimera dal punto di vista genetico, cioè una entità composta da parti riunite insieme da fonti disparate? In quale stadio (o in quali stadi) dell'evoluzione questo presunto montaggio avrebbe avuto luogo? E qual era, poi, la natura degli organismi che hanno fornito i vari geni e le diverse strutture?

I biologi hanno sempre avuto la tendenza a pensare che la cellula eucariota fosse formata dall'associazione di cellule procariote che avevano raggiunto la piena evoluzione; essi ammettono che questa associazione abbia creato un tipo di cellula «superiore», la cellula eucariota. (La parola procariota - «prima del nucleo» - sottintende proprio questo.) L'interrogativo che ci si potrebbe utilmente porre adesso è se gli eventi evolutivi che diedero alla cellula eucariota il suo carattere molecolare fondamentale siano stati realmente di questa natura. La cellula eucariota sembra essere una «chimera» in senso biologico già a un livello molto fondamentale e perfino il ribosoma eucariota risulterebbe tale, essendo i suoi RNA di una origine diversa da quella di perlomeno una delle sue proteine. Se questa è una interpretazione corretta dei dati (e le ricerche future dovranno risolvere questo punto), la cellula eucariota potrebbe essere un tipo di entità diverso da quello che si ritiene oggi. Potrebbe essere stata una chimera ancora prima di aver raggiunto uno stadio di complessità paragonabile a quello dei procarioti odierni; potrebbe essere stata una chimera fin da quando emerse dalla condizione di progenote. Invece di essere una forma «superiore», progredita, la cellula eucariota potrebbe rappresentare una regressione alla dinamica evolutiva del suo lontano antenato, il progenote.

Forse l'aspetto più stimolante riguardo alle recenti scoperte sulla filogenesi molecolare è che mostrano quante informazioni sui primissimi stadi dell'evoluzione siano racchiuse nella cellula. Non è più necessario basarsi esclusivamente sulle congetture per spiegare le origini della vita. Ci si è abituati a pensare agli ultimi decenni di questo secolo come a un periodo della biologia in cui la «ingegneria genetica» ha permesso di realizzare sorprendenti progressi in campo medico e industriale. È necessario però riconoscere anche che la biologia si trova oggi sulla soglia di una rivoluzione meno sconvolgente, in cui l'uomo arriverà a capire le radici di ogni forma di vita, e, pertanto, acquisirà una comprensione più profonda del processo evolutivo.

Il rischio di frane nelle Alpi

Nelle valli di Courmayeur le frane della Brenva e del Triolet, avvenute nel recente passato, dimostrano come tali fenomeni rappresentino un potenziale pericolo che non si deve sottovalutare

di Giuseppe Orombelli e Stephen C. Porter

L'aumento della popolazione, il progresso tecnologico, l'industrializzazione e il massiccio aumento dei mezzi di trasporto hanno profondamente modificato in questi ultimi decenni i rapporti tra l'uomo e la natura. In Italia, in particolare, negli ultimi trent'anni la popolazione è cresciuta di circa 10 milioni di abitanti. Il numero delle abitazioni è passato da circa 10 ad oltre 17 milioni. I territori urbanizzati si sono conseguentemente allargati a scapito di quelli destinati a uso agricolo o comunque alle aree verdi. Il numero degli autoveicoli in circolazione è cresciuto da meno di un milione a oltre 17 milioni (circa un autoveicolo ogni tre abitanti). Il turismo è divenuto fenomeno di massa.

È comprensibile quindi come nuovi problemi si siano posti nell'uso del territorio o più spesso si siano fortemente aggravati quelli già noti. In particolare vogliamo qui ricordare i problemi ambientali legati agli aspetti fisici del territorio, comunemente indicati come geologici, in senso lato. Con il termine di «rischio geologico» si intende oggi alludere a tutti i possibili effetti dannosi che la natura geologica di un territorio può avere nei riguardi delle attività umane e della vita umana stessa. Sono gli aspetti geologico-dinamici quelli che potenzialmente costituiscono rischio geologico, legati ai processi interni della crosta terrestre, quali i terremoti o l'attività vulcanica, o ai processi esterni, quali alluvioni, movimenti del suolo, frane, valanghe e così via. Tali processi hanno sempre operato in tutto il corso dei tempi storici e geologici, ma attualmente, con l'aumento della popolazione e con l'addensarsi di essa in centri urbani, sono divenuti potenzialmente più pericolosi, in grado di causare danni sempre più gravi.

L'ambiente alpino è indubbiamente tra quelli più dinamici dal punto di vista geologico ed è particolarmente esposto all'azione dei processi esterni. In particolare, malgrado una falsa impressione di imponente staticità delle sue masse montuose e delle sue pareti rocciose, esso è ampiamente esposto al «rischio di frana», non fosse altro che per l'accentuata «energia del rilievo», cioè per i forti dislivelli tra fondovalle e sommità dei versanti, creste e cime, e per le pendenze spesso molto elevate.

Inoltre l'ambiente di alta montagna, sino a qualche decennio fa praticamente disabitato o soltanto sporadicamente frequentato durante l'estate, è divenuto sede di un afflusso turistico sempre più imponente e non soltanto nei mesi estivi, ma anche in quelli invernali e primaverili, per il richiamo esercitato dallo sci. Impianti sciistici, attrezzature turistiche, abitazioni, campeggi si sono insediati sempre più diffusamente e sempre più in alto nelle nostre montagne. Grazie alla facilità di accesso e di riscaldamento si vive oggi con la montagna in un rapporto di tranquilla confidenza, come certamente non si era mai verificato nei secoli passati. Eppure non mancano testimonianze storiche di passate vicende in cui la montagna è stata causa di distruzione e rovina.

Restando soltanto nel campo delle grandi frane che hanno causato vittime e danni ingenti, seppellendo talora interi villaggi, lunghi elenchi sono stati compilati per le diverse regioni alpine. Per ricordare soltanto alcuni esempi relativi al versante meridionale delle Alpi negli ultimi cinque secoli, si può citare la grande frana (oltre 15 milioni di metri cubi) che nel 1512 sbarrò la Val Blenio, nel Canton Ticino, creando un lago; lo svuotamento improvviso di quest'ultimo, due anni dopo, causò la morte di 600 persone. Nel 1618 una frana catastrofica (oltre 3 milioni di metri cubi) distrusse il paese di Piuro, presso Chiavenna, causando un migliaio di morti. Nel 1648 una frana distrusse il villaggio di Fano presso Brentonico (Trento) e nel 1692 una grande frana (30 milioni di metri cubi) seppellì il villaggio di Borta, nell'alta valle del Tagliamento, creando per sbarramento un lago della lunghezza di quasi 7 chilometri, estintosi successivamente per erosione e per interramento. La frana di Alleghe (20 milioni di metri cubi), crollata nel 1771, distrusse tre piccoli villaggi e diede luogo anch'essa a un lago di sbarramento, tutt'ora esistente, anche se di dimensioni ridotte rispetto a quelle iniziali. Nel 1868 una frana presso Bodio (Canton Ticino) distrusse un piccolo villaggio. Infine va citata la recente disastrosa frana del Vajont (1963, volume stimato 270--300 milioni di metri cubi), il cui tragico ricordo pesa ancora sul nostro paese.

Altre frane, talora ancor più grandi, si sono verificate in tempi più antichi: così la frana dei Lavini di Marco (150-200 milioni di metri cubi) e quella di Dro (da 200 a 700 milioni di metri cubi) nelle valli dell'Adige e del Sarca, si sarebbero staccate rispettivamente nell'alto Medioevo e in epoca romana, ma la loro datazione è alquanto incerta e controversa e inoltre si tratta di frane complesse, ripetutesi più volte in tempi successivi. L'elenco potrebbe proseguire a lungo: un panorama completo si può trovare nelle opere di A. Heim (1932) e di G. Abele (1974).

Se il numero delle frane documentateci dalla storia è grande, ancor più grande è il numero delle frane testimoniateci da evidenze geologiche. Mentre le prime sono soltanto quelle di grandi dimensioni che colpirono centri abitati, in particolare durante l'evo moderno (da quando cioè la documentazione storica relativa alle alte vallate alpine diventa più consistente), le seconde registrano fedelmente il fenomeno nell'intero areale alpino e in un intervallo di tempo assai più ampio, indicativamente gli ultimi 10 000 anni circa, cioè l'intero Olocene, l'intervallo di tempo successivo al ritiro dei ghiacciai dell'ultima avanzata tardo-würmiana.

Un primo passo nella valutazione del rischio di frana consiste nel conoscere dove, quando e con quali modalità si sono verificate grandi frane nel recente passato geologico, prima di procedere a più dettagliati studi geologico-tecnici sulla stabilità delle masse rocciose. È importante valutare, per una stessa vallata, il tempo di ritorno delle frane in funzione della loro entità (volume, superficie, distanza percorsa), almeno per una previsione statistica del fenomeno (distribuzione areale, dimensioni e frequenza). L'identificazione delle aree di distacco e, più ancora, degli accumuli di frana, della loro estensione e della loro età è quindi il primo compito assegnato al geologo.

Le nicchie di distacco delle frane appena formatesi sono individuabili facilmente in base alla loro forma e al colore delle nuove superfici esposte, ma spesso questi aspetti vengono rapidamente modificati e cancellati dai processi di alterazione ed erosione. Anche l'identificazione e la delimitazione degli accumuli di frana è, talora, un compito non semplice, perché possono essere scambiati per accumuli di altra natura, per esempio glaciali, nelle aree in cui vi sono o vi furono ghiacciai. Ma indubbiamente è la datazione delle frane il compito più difficile. In questo articolo si vogliono descrivere due casi di frane verificatesi in tempi storici a noi vicini nelle valli di Courmayeur, per indicare una possibile via di studio, nella quale dati storici, geologici, geomorfologici e botanici vengono simultaneamente utilizzati per l'identificazione e la datazione di questi fenomeni.

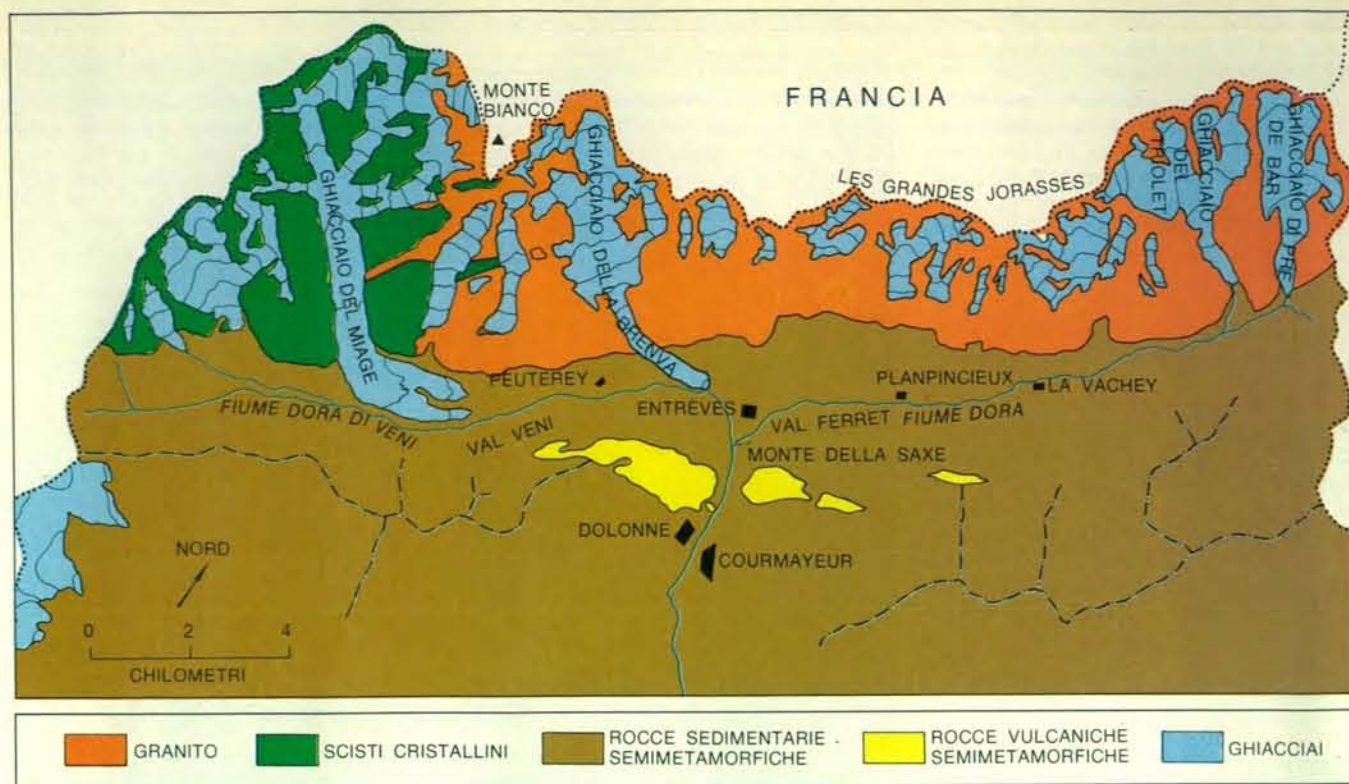
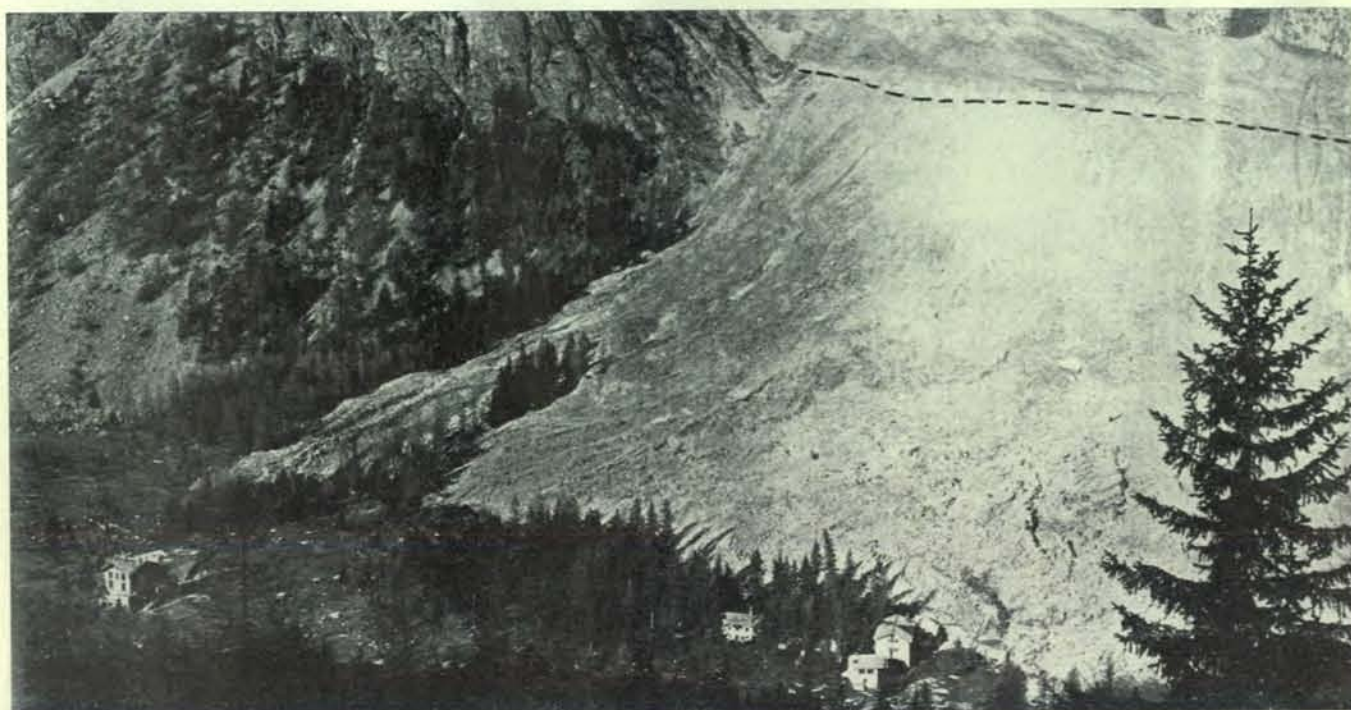
Le valli di Courmayeur

Il massiccio del Monte Bianco, nelle Alpi Occidentali, si presenta dal versante italiano come una imponente e continua barriera di alte montagne, orientata da nord-est a sud-ovest, con pareti rocciose, creste e guglie aguzze e numerosi ghiacciai, per lo più annidati in circhi e valloni e affaccianti sul fondovalle. Soltanto i ghiacciai di dimensioni maggiori superano con serraccate le soglie dei circhi e il dislivello che li separa dal fondovalle, riversandosi in esso come fiumi di ghiaccio.

Ai piedi del gruppo del Monte Bianco, sul versante italiano, si estendono due valli, anch'esse orientate da nord-est a sud-ovest, la Val Ferret e la Val Veni, confluenti nei pressi di Entrèves in un'unica valle, orientata da nord nord-ovest a sud sud-est. Le due valli Ferret e Veni hanno un profilo trasversale accentuatamente asimmetrico. Il loro fianco settentrionale, quello del Monte Bianco, è estremamente ripido ed elevato. La linea di cresta è quasi interamente sopra i 3000 metri e per ampi tratti sopra i 4000. Dislivelli di 2-3000 metri e oltre sono superati nello spazio di pochi chilometri, con pendenze medie, da cresta a valle, localmente sfioranti il 100 per cento. Frequentissime sono le pareti rocciose subverticali, con dislivelli di parecchie centinaia di metri. Il versante meridionale, opposto al gruppo del Bianco, è invece assai più basso, raggiungendo solo localmente i 3000 metri, i dislivelli da cresta a valle sono inferiori a 2000 metri e la pendenza media è in generale inferiore al 50 per cento.

L'orientazione e l'asimmetria del profilo trasversale delle valli di Courmayeur sono dovute all'assetto geologico del territorio, contraddistinto dall'accostamento, secondo un allineamento nord-est sud-ovest di unità strutturali differenti. In particolare il gruppo del Bianco è costituito pressoché interamente da granito e da scisti cristallini (migmatiti e gneiss) mentre il versante meridionale

La Val Ferret nei pressi di Courmayeur. È evidente l'asimmetria dei due versanti: quello destro, che corrisponde al versante italiano del gruppo del Bianco, è estremamente ripido ed elevato; quello sinistro è invece assai più basso e ha una debole pendenza.

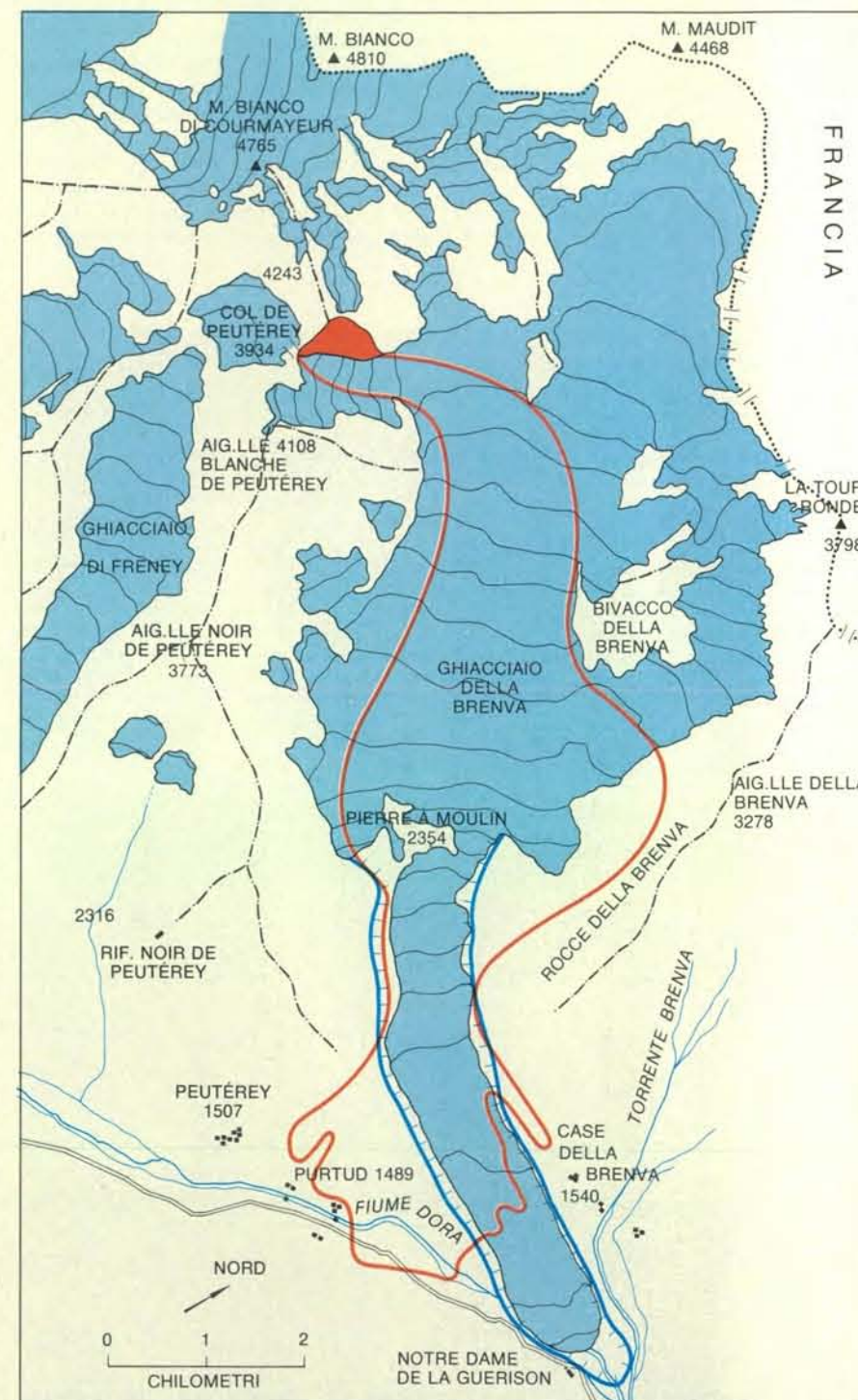
Anche la Val Veni presenta le stesse diversità tra i due versanti sia per quanto riguarda i dislivelli, sia per quanto riguarda le pendenze. Sullo sfondo della fotografia si possono osservare i due grandi lobi a tenaglia del ghiacciaio del Miage interamente coperto di detriti.

Il versante italiano del massiccio del Monte Bianco e le valli di Courmayeur. Come mostra la cartina il massiccio del Monte Bianco è quasi interamente costituito da granito e da scisti cristallini (migmatiti e gneiss), mentre il versante meridionale delle valli di Courmayeur è costituito da rocce sedimentarie e vulcaniche semimetamorfiche. Alla asimmetria di costituzione geologica corrisponde una netta asimmetria morfologica. Il versante meridionale del Bianco è assai ripido ed elevato, con dislivelli di 2-3000 metri e pendenza media che localmente sfiora il 100 per cento, come nel settore mediano della Val Ferret. Numerosi ghiacciai occupano le valli secondarie e i maggiori di essi scendono fino al fondovalle principale. Frequentissime sono le pareti rocciose subverticali e le guglie di roccia. Il fianco meridionale delle valli di Courmayeur presenta invece dislivelli inferiori a 2000 metri e la pendenza media è inferiore al 50 per cento. È praticamente privo di ghiacciai.

La frana del novembre 1920 alla Brenva in una fotografia dell'epoca, pubblicata sul n. 50 dell'«Illustrazione italiana». L'intero versante esterno della morena destra del ghiacciaio - prima ricoperto da un fitto bosco - appare spazzato dalla frana, i cui materiali, roccia e ghiaccio, giungono sino alle case di Purtud. Si notino in basso a sinistra i massi, probabilmente deposti da una precedente frana. La linea tratteggiata indica la cresta della morena, più bassa della superficie del ghiacciaio. Questo nel 1920 era in una fase di aumento di volume; a seguito della frana e della copertura detritica che ne rallentò l'ablazione, la Brenva prolungò in maniera anomala tale fase di aumento sino al 1941, avanzando la sua fronte sin quasi a raggiungere la posizione di massima estensione toccata nel 1818.

delle valli è costituito prevalentemente da rocce sedimentarie semimetamorfiche (argilloscisti, calcari, gessi, calcescisti, ecc.). Come si è detto il versante italiano del Monte Bianco è costellato da ghiacciai. I maggiori sono, in Val Veni, i ghiacciai della Lex Blanche, del Miage, della Brenva e, in Val Ferret, i ghiacciai del Triolet e di Pré de Bar.

Le frane della Brenva

Il giorno 14 novembre 1920, alle ore 14 e 30 circa, una massa di roccia e ghiaccio si staccò dai contrafforti orientali del Monte Bianco di Courmayeur, poco sotto la quota 4243 (fianco est sud-est del Grand Pilier d'Angle), precipitando sul sottostante ghiacciaio della Brenva. Testimoni oculari, richiamati dal rumore, fecero la seguente descrizione del fenomeno. Dapprima fu vista la massa di roccia slittare lungo la parete, lasciandosi dietro come due scie bianche di polvere, quindi abbattersi sul sottostante canalone che porta al Col de Peutérey e ribaltarsi interamente, mostrando la superficie di distacco, di colore più chiaro («voltarsi come questo cappello che mostra la fodera»). Frantumatasi la massa, si sviluppò una densa nuvola di polvere bianca, dai contorni ben definiti, «simile a un fiocco di lana», che prese a scendere lungo il ghiacciaio della Brenva a grande velocità. Passando sopra la rottura di pendio e la serraccata della Pierre à Moulin si videro pietre e massi schizzare fuori dalla nuvola come fossero stati sparati da una mina. Giunta sulla parte inferiore del ghiacciao la frana-valanga si suddivise in due flussi distinti che traboccarono sui due lati del ghiacciaio oltre le morene laterali, raggiungendo il vallone delle Case della Brenva sulla sinistra e il Piano del Peutérey sulla destra. Dietro il passaggio della «nuvola» si levò una fitta nebbia che per parecchio tempo occultò la vista dell'alto bacino del ghiacciaio. Il dislivello percorso fu di oltre 2000 metri, la distanza di circa 5 chilometri, il tempo impiegato di 2-3 minuti, la velocità, quindi, di oltre 100 chilometri all'ora.

Cinque giorni più tardi, il 19 novembre 1920, tra le 14 e le 16 e 45 si ebbero quattro nuovi crolli consecutivi, di cui l'ultimo fu il più catastrofico di tutti. Il materiale di frana, roccia e ghiaccio, traboccò per un fronte di circa 1 chilometro oltre la morena destra, rovesciandosi sul versante esterno della stessa. Si videro dapprima massi di pietra e blocchi di ghiaccio isolati «balzare dalla nebbia sovrastante il ghiacciaio, saltando il ciglio della morena in mezzo alla foresta, indi ottenebrarsi il cielo completamente». L'accumulo di frana raggiunse il piano del Peutérey, fermandosi a ridosso delle case di Purtud e risalì in parte sul versante opposto della Val Veni, sbarrando il corso della Dora. Si formò un lago che ebbe però vita assai breve, anche per il pronto intervento di sfondamento effettuato dagli abitanti. Al termine dei crolli la nicchia di distacco si estendeva a tutta la parete sud-est dello sperone di quota 4243 (Gran Pilier d'Angle), dalla sua sommità presso il colletto del Peutérey alla base del canalone, per un dislivello di circa 6-700 metri. Il volume di roccia crollato fu stimato a circa 6 milioni di metri cubi. Il colle del Peutérey era quotato prima della frana 3987 metri, attualmente, invece, è quotato 3934 metri.

Mentre il percorso della frana-valanga del 14 novembre seguì la linea mediana del ghiacciaio della Brenva, passando a destra della Pierre à Moulin e biforcandosi ai piedi della serraccata nei due lobi già descritti, le successive frane-valanghe del 19 attraversarono obliquamente il

Il versante meridionale del Bianco e il ghiacciaio della Brenva. L'area rossa sul fianco est sud-est del Gran Pilier d'Angle (quota 4243) indica la zona di distacco delle frane del 14 e 19 novembre 1920. La linea in rosso delimita l'area percorsa dalle frane. Nel complesso queste si scaricarono dapprima nel canalone che porta al Col de Peuterey, quindi sull'alto bacino della Brenva, risalendo in parte sul versante opposto delle Rocce della Brenva e abbattendosi quindi sulla parte inferiore del ghiacciaio, ove traboccarono oltre le morene laterali sui due lati, particolarmente su quello destro. Il dislivello massimo perorso fu di oltre 2000 metri, la distanza massima di circa 5 chilometri, la velocità di oltre 100 chilometri all'ora, il volume stimato di almeno 7 milioni di metri cubi. La linea in blu indica la massima estensione raggiunta dalla Brenva, all'inizio del secolo XIX (1818).

bacino superiore della Brenva, risalendo per circa 50 metri sul fianco ovest delle Rocce della Brenva (Mont de la Brenva), per rimbalzare poi ancora obliquamente verso il margine destro del tratto terminale del ghiacciaio, traboccando oltre la morena. L'accumulo di frana (roccia e ghiaccio) sopra il solo tratto terminale della Brenva e sui fianchi esterni delle morene laterali fu stimato a 6-7 milioni di metri cubi. Esso risultava caratterizzato da una struttura a lobi, con creste allungate, alte anche 10 metri, e archi concentrici trasversali. La caduta delle frane fu accompagnata da una nuvola di pulviscolo trasformatosi poi in un fitto nebbione. Questo risultò costituito oltre che da polvere di ghiaccio, da polvere di roccia triturata, delle dimensioni del limo e della sabbia molto fine, che si distribuì sul territorio circostante, anche a monte della frana e sul versante sud della Val Veni fino a 1800 metri di quota circa, formando uno straterello dello spessore di circa 2-3 millimetri.

Le frane del novembre 1920 alla Brenva non causarono fortunatamente vittime né danni diretti alle case, ma distrussero circa 50 ettari di bosco. A seguito delle frane del 1920 il ghiacciaio della Brenva conobbe una fase di avanzata eccezionale, assolutamente anomala rispetto agli altri ghiacciai del gruppo del Monte Bianco, a causa della copertura detritica e della conseguente diminuita ablazione (come si è riscontrato anche su altri ghiacciai colpiti da grandi frane). Tale fase si protrasse fino al 1941 e portò la fronte del ghiacciaio a raggiungere quasi le posizioni assunte nelle avanzate dell'inizio e della metà del secolo scorso, quando il ghiacciaio ebbe la massima estensione dalla fine dell'ultima glaciazione, cioè negli ultimi 10 000 anni circa.

A distanza di 60 anni cosa rimane della frana della Brenva, cosa si può vedere sul luogo del disastro? (Per semplicità, anche se si è trattato di distinti successivi crolli, faremo d'ora innanzi riferimento a un solo evento.) La grande ferita creata nel manto vegetale è stata gradualmente cicatrizzata. La foresta di conifere (larici, abeti rossi) si è nuovamente estesa a coprire l'accumulo di frana. Tuttavia proprio la vegetazione ci offre oggi una evidente testimonianza dell'evento del 1920. Infatti un limite netto di vegetazione sottolinea il margine estremo raggiunto dalla frana. Lungo questo limite vengono a contatto, sul lato esterno, gli alberi «pre-frana», alti e ben sviluppati, sul lato interno, gli alberi «post-frana», di taglia nettamente inferiore. I primi sono di età più che centenaria, come risulta dal conteggio degli anelli (o cerchie) annuali di accrescimento, effettuato su sottili «carote» estratte dai loro tronchi. In taluni di essi la data stessa della frana è rimasta registrata da un improvviso aumento dello spessore degli anelli di accrescimento, realizzatosi a partire dall'anello formatosi nel 1921, quando cioè l'albero, in precedenza circondato da altri individui al centro del bosco, si trovò improvvisamente posto in condizioni migliori di illuminazione, al margine del bosco. Nei larici e negli abeti rossi cresciuti sopra l'accumulo di frana il numero degli anelli di accrescimento annuale, a 0,5 metri dal suolo, non supera i 40-45 (i conteggi in questo articolo sono riferiti all'estate del 1979). Questo dato indica anche che, dal momento della frana, trascorsero circa 15-20 anni prima che germogliassero nuovi larici e abeti rossi e questi raggiungessero un'altezza di circa mezzo metro.

Un altro aspetto della vegetazione ci può aiutare oggi a riconoscere e a delimitare esattamente l'estensione della frana del 1920. Si tratta della copertura lichenica. Sui massi esterni e più antichi della frana i licheni di vario tipo si sono per lungo tempo sviluppati e sono quindi abbondanti e di grandi dimensioni, mentre sui massi della frana i licheni sono più scarsi e più piccoli.

La frana della Brenva del 1920 è ovviamente anche riconoscibile nei suoi aspetti geologici e geomorfologici. L'accumulo di frana sul versante esterno della grande morena destra del ghiacciaio è costituito da un caos di grossi massi monolitologici, cioè tutti costituiti da «protogino», il granito a struttura porfirica (con grossi cristalli di feldspato) che costituisce quasi interamente il massiccio del Monte Bianco (salvo la sua porzione più occidentale). I massi, generalmente spigolosi o con contorni subarrotondati, hanno dimensioni fino a parecchie decine di metri cubi. Tra essi sono frammisti blocchi e altri frammenti rocciosi di dimensioni via via minori, da quelle di un ciottolo a quelle della sabbia e dei limi.

Il ghiacciaio della Brenva visto dal versante opposto della Val Veni, dai pressi del rifugio CAI Uget. Si riconoscono la vetta del Monte Bianco di Courmayeur (m. 4765) e il Gran Pilier d'Angle, dal cui fianco meridionale si staccò la frana del 1920. Sono visibili la «Pierre à moulin» e cioè il tratto di roccia scoperto nel mezzo della serraccata, una parte del tratto inferiore del ghiacciaio ricoperto da detriti e l'attacco della grande morena laterale destra.

I massi dell'accumulo di frana si riconoscono, oltre che per la scarsa copertura lichenica, per l'aspetto superficiale di roccia inalterata, da poco tempo esposta, cioè, all'azione degli agenti atmosferici, a differenza di altri massi più antichi osservabili, per esempio, nei pressi di Purtud o, nelle vicinanze, sul versante nord della Val Veni, caratterizzati da una superficie più alterata, scabra, con cristalli in rilievo ed estesa copertura lichenica.

L'accumulo di frana mostra un margine netto, con uno spessore di pochi metri rispetto alla superficie topografica su cui si appoggia. La morfologia dell'accumulo di frana è tipicamente accidentata, caratterizzata cioè da un irregolare alternarsi di depressioni e rilievi, con un dislivello locale dell'ordine anche di parecchi metri. L'attraversamento del bosco sviluppatosi sopra l'accumulo di frana è quindi estremamente disagevole e faticoso.

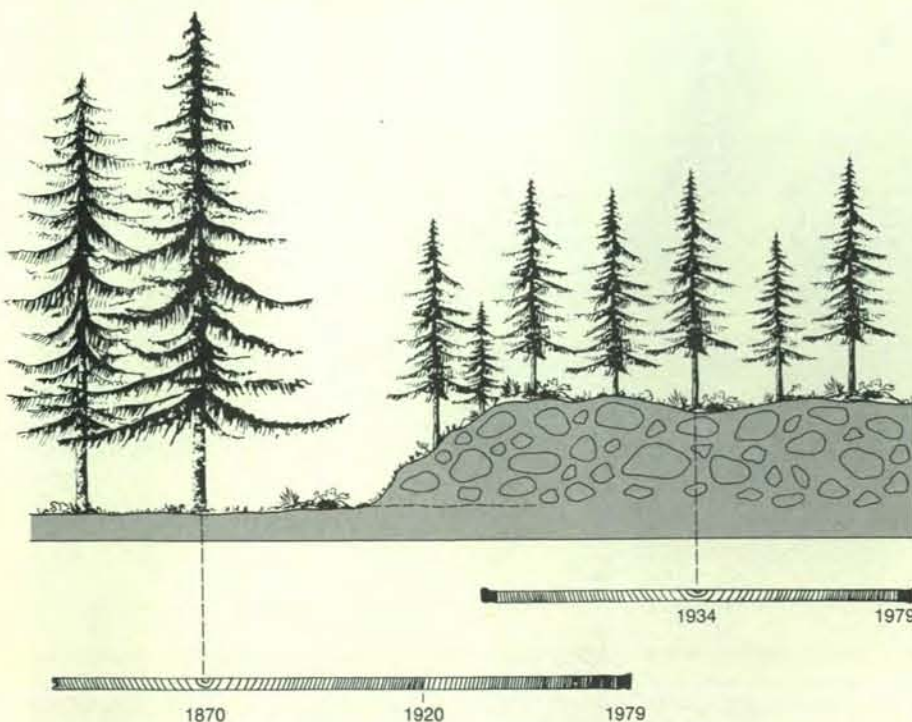
Malgrado quanto è stato detto sin qui, dato che l'accumulo della frana della Brenva del 1920 si è deposto sopra il fianco esterno della morena destra del ghiacciaio, esso potrebbe essere attribuito, in assenza delle testimonianze dirette dell'evento e delle tracce ancora visibili nella vegetazione, non all'azione di una frana ma soltanto a quella del ghiacciaio. In ambiente alpino, infatti, un accumulo di frana può non differire, a un primo sguardo, da un accumulo morenico, particolarmente se la litologia dei rilievi circostanti è uniforme.

È pensabile, quindi, che se la frana della Brenva si fosse prodotta in epoca più lontana, senza che di essa fosse stata lasciata testimonianza storica e la vegetazione avesse avuto il tempo di cancellarne interamente le tracce, essa avrebbe potuto facilmente passare inosservata. In altre parole, non ci si sarebbe resi conto di un evento naturale di dimensioni eccezionali, indicativo di un potenziale rischio geologico.

La frana del Triolet: documenti storici

Un evento simile a quello della Brenva si verificò circa due secoli prima in alta Val Ferret. Di esso si posseggono testimonianze storiche. Malgrado ciò la realtà geologica dell'evento andò gradualmente oscurandosi al punto che gli studiosi che, anche recentemente, si sono occupati di

In questa fotografia, eseguita nel 1975 dalla Compagnia Generale Ripreseaeree di Parma per conto della Regione autonoma Valle d'Aosta (concessione SMA n. 78 dell'11 febbraio 1981), è visibile la lingua terminale del ghiacciaio della Brenva. L'area devastata dalla frana del 1920 sul versante esterno della morena destra è riconoscibile per uno sviluppo diverso della vegetazione. Gli alberi sviluppatisi sulla frana, essenzialmente larici e abeti rossi, sono più giovani di quelli che li circondano, costituiti dalle medesime essenze, ma di colore verde più scuro.

Osservando la vegetazione è possibile identificare con precisione l'estensione raggiunta dalla frana della Brenva. Sono a contatto abeti rossi e larici precedenti alla frana, di grande taglia e di età più che centenaria, con le stesse specie sviluppatesi sulla frana, di taglia minore e di 40-45 anni di età o più giovani. Il grosso masso di frana in primo piano ha un contorno angolare che indica che esso ha subito un limitato trasporto. La superficie chiara e non alterata e la scarsa copertura lichenica dimostrano inoltre che la superficie rocciosa è stata esposta di recente. Un altro grosso masso dello stesso tipo è visibile in distanza, sul margine della strada.

Schizzo della situazione al margine della frana della Brenva del 1920. Le sottili carote estratte mediante trivellina dai tronchi degli alberi rivelano per quelli cresciuti sull'accumulo di frana un'età massima di 45 anni, mentre quelli precedenti alla frana sono di età secolare. Talora questi possono rivelare anche l'età esatta (l'anno) della frana, quando si osservi, come nel caso illustrato, un improvviso aumento dello spessore degli anelli di accrescimento a partire dall'anno successivo all'evento franoso. Il fatto si spiega ricordando che l'albero in questione prima della frana cresceva in mezzo al bosco e a causa della frana è venuto improvvisamente a trovarsi al suo margine, in condizioni di migliore illuminazione e quindi di più rapido accrescimento. Questo genere di osservazioni e altre simili consentono alla dendrogeomorfologia di datare con la approssimazione di un anno frane o altri eventi geomorfici, quali alluvioni, valanghe, avanzate di ghiacciai, ecc. che hanno bruscamente modificato le condizioni di crescita delle piante.

questa zona non hanno collegato tra loro documentazione storica e testimonianze geologiche. L'evento rimase confinato negli scritti, mentre sul luogo si perse l'esatta percezione dello stesso.

Nell'archivio della parrocchia di Courmayeur si conserva una pergamena messaci gentilmente a disposizione da don Cirillo Perron e riprodotta nella pagina a fronte. Su di essa, in un latino approssimativo e sgrammaticato, si legge: «Mentre era in corso una gran lite per un piccolo pezzo di terra tra - i proprietari dei pascoli di Trioley e di Ameiron, la vetta di un monte - sul lato sinistro del ghiacciaio detto Trioley, la notte del giorno dodici settembre - dell'anno 1717 improvvisamente cadde, e le sue rocce, le acque e i ghiacci insieme - riuniti con grande impeto precipitarono sopra i suddetti pascoli - o alpeggi, cosicché coperti sotto quell'abisso tutte - le mobilia, centoventi buoi o vacche, formaggi, uomini - in numero di sette morirono all'istante e credo che - gli uccelli che ivi nidificavano non riuscirono a sfuggire, ed ora vedete - quegli alpeggi e quella piana in questo stato miserevole, e ciò - a causa dei nostri peccati. Così attesto perché vidi - Michele Giuseppe Pennard - notaio e copista in - Courmayeur.» (Traduzione di M. Pozzi).

Il documento non è datato, ma è da ritenersi immediatamente successivo all'evento. Pochi anni più tardi lo storico J. B. De Tillier, nella sua opera manoscritta nota come *Historique de la Vallée d'Aoste*, della quale esistono diverse edizioni datate dal 1721 al 1740, così riferisce l'evento (traduzione dal testo francese datato 1737 e pubblicato nel 1966): «È sul fondo della medesima valle (la Val Ferret), che ha circa sette-ottocento passi ordinari di larghezza pressoché in tutto il suo corso, che la notte dal 7 all'8 di settembre dell'anno 1717 accadde uno strano disastro: un alto picco di roccia, carico sul suo dorso d'un pesante fardello di ghiaccio, essendosi staccato in un sol colpo, rovinò sopra un altro grande ghiacciaio esistente in uno stretto vallone a sinistra della sopraddetta valle, il quale, essendosi trovato vuoto di sotto, si ruppe in numerosi pezzi. Il ghiaccio così rotto, il crollo e il gran peso dell'altro, unitamente alle acque frammischiate, avendo dato movimento e servito di veicolo, fecero avanzare una infinità di pietre e grossi roccioni staccati dalla sopraddetta montagna, che urtandosi nel loro cammino e rotolamento mescolati tra il ghiaccio, fecero un disordine così spaventoso da sorpassare l'immaginazione, riempiendo tutta la larghezza della valle sopraddetta per circa una lega di estensione, di pietre, ghiaia e grossi massi, facendone salire sulle pareti laterali di così grandi quanto delle case, e coprendo interamente tre alpeggi con le loro abitazioni, praterie, pascoli, persone, bestiame e frutti di tutta la campagna, senza che ne resti alcun segno, né pure che si possa positivamente riconoscere il luogo ove le abitazioni erano situate.»

A parte la discordanza circa il giorno dell'evento, dai testi di Pennard e di De Tillier si può ricavare una precisa descrizione del fenomeno, ricca di dettagli, che ci consente, tra l'altro, di ubicare l'area di distacco della frana nel versante occidentale dell'alto recinto montuoso del ghiacciaio del Triolet. La frana, rovesciatasi sopra il ghiacciaio stesso, raggiunse la valle principale (Val Ferret) estendendosi e distruggendo ogni cosa per una distanza di qualche chilometro.

Nella seconda metà del secolo XVIII ha inizio l'esplorazione naturalistica e geologica delle Alpi. Di essa il pioniere è Orazio Benedetto De Saussure. Nell'estate del 1781 De Saussure percorre la Val Ferret e osserva il ghiacciaio del Triolet «coperto dalle macerie di un'alta montagna di granito», e annota: « Circa 60 anni fa, dall'alto di questa montagna, si verificò una grande frana, che seppellì in una notte le abitazioni che erano ai piedi di questo ghiacciaio, con gli animali e i pastori. Da allora queste frane sono sempre continuate.»

Sessant'anni più tardi J. D. Forbes, nel 1841, percorre e descrive la Val Ferret e osserva, a proposito del ghiacciaio del Triolet, come esso discenda «da una cima chiamata dalle guide di Chamouni Montagna delle frane». E aggiunge: «L'evento al quale il nome fa riferimento ebbe luogo, io credo, nel 1728, sebbene io non sia riuscito ad ottenere a Courmayeur alcun autentico documento riguardo ad esso. Secondo un opuscolo stampato che mi fu mostrato, la valanga, o improvvisa discesa dell'intero ghiacciaio, ebbe luogo nella notte del 15-16 agosto di quell'anno, e completamente annientò le baite di Pré de Bar, che erano situate esattamente di fronte ad esso, distruggendo naturalmente gli abitanti ed il bestiame.»

Si noti come, con il trascorrere degli anni, data, luogo, estensione e modalità dell'evento vadano gradualmente oscurandosi.

I primi studi geologici e geomorfologici del versante meridionale del Bianco datano dalla metà del secolo scorso. Al fondo della Val Ferret vengono osservati alcuni depositi detritici molto grossolani che vengono unanimemente attribuiti a una passata maggiore estensione dei ghiacciai del Triolet e di Pré de Bar. Tale interpretazione viene mantenuta fino ai nostri giorni. Vi è solo discordanza per quanto riguarda l'età di tale espansione, da taluni ritenuta di età storica, da altri e più recentemente, di età preistorica. Il testo del Pennard, tornato nel frattempo alla luce, viene riprodotto e citato come curiosità da numerosi naturalisti e geologi, ma ad esso non viene dato gran credito o prestata sufficiente attenzione.

Questa pergamena conservata nell'archivio della parrocchia di Courmayeur, firmata dal notaio Pennard, documenta la frana del Triolet del 1717. Il testo latino, sgrammaticato e approssimativo, può essere così letto: «Dum lis ingens ventilabatur de modico terrarum inter - possessores montium du triolley et d'Ameiron, mons excelsissimus - versus sinistris glaciei dicti trioley, nocte diei duo decimi septembris - anni 1717 subito cecidit, et saxa eius, aquae et glacies simul - conglobata cum magno impetu irruerunt super dictos montes - sive alpes, ita ut cooperti subtus (?) ipsis abissis omnia - mobillia, centum et vingenti boves seu vaccarum, casej, homines - ad numerum septem qui instanti perierint et credo quod - aves qui ibi nidicafabant non effugerint, et hunc videtis-illos alpes et illam planam in hoc miserrimo statu, et hoc, propter scelera nostra. Ita atestor quia vidi - Michael Joseph Pennard - notarius et scriba Curiae - Majoris.» La traduzione è riportata a pagina 74. La frana descritta da Pennard non fu riconosciuta sul terreno nei secoli successivi dai naturalisti e geologi che studiarono il versante meridionale del Bianco. È questo un esempio di come i documenti storici possano essere di grande importanza nella ricostruzione delle vicende geologiche recenti.

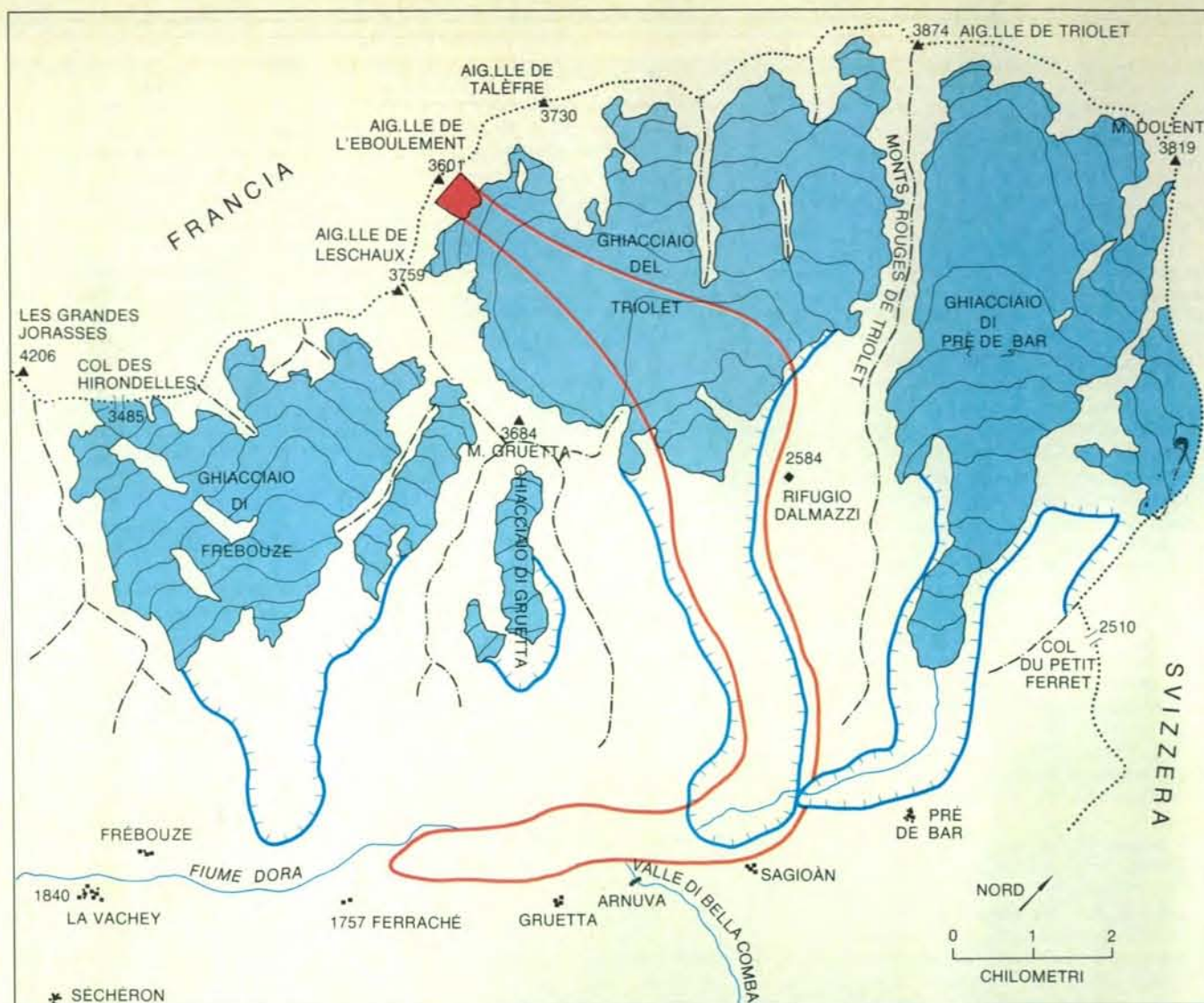
La frana del Triolet: prove geologiche e geomorfologiche

Dal momento che gradualmente era andata perduta ogni consapevolezza geologica della frana del Triolet, bisognava nuovamente riscoprirla sul terreno. Le osservazioni fatte sullo stato attuale della frana della Brenva potevano costituire una traccia dei metodi da seguire a questo scopo: identificare l'accumulo di frana sulla base dei suoi aspetti litologici e geormofologici, quindi di giungere a una sua datazione.

Percorrendo la Val Ferret, oltre La Vachey, superato il torrente che scende dalla valle di Malatrà, si giunge nella piana acquitrinosa di Ferraché. Al fondo di questa si osserva un accumulo di grossi blocchi, il cui fronte arcuato si staglia nettamente, elevandosi di qualche metro sopra la pianura. I blocchi che lo costituiscono possono raggiungere un volume di parecchie centinaia di metri cubi (il più grande misura 20×12×4 metri) e sono esclusivamente granitici. Questo accumulo di massi è allungato lungo il fondovalle e lo si può seguire, sia pure discontinuamente, per circa 2 chilometri sino oltre Arnuva. La sua larghezza è in media di circa 500 metri. In pianta ha un andamento leggermente serpeggiante, appoggiandosi ora a un lato ora a quello opposto della valle. La sua superficie è caratterizzata da una morfologia accidentata, con depressioni, colmate in parte da alluvioni, e rilievi irregolarmente distribuiti, con dislivelli che localmente possono raggiungere i 7 metri. Sono inoltre presenti rilievi stretti e allungati longitudinalmente, sopraelevati rispetto alla superficie circostante di qualche metro. Nel tratto tra Arnuva e le case di Gruetta l'accumulo di blocchi è in gran parte sepolto da depositi alluvionali portati dal torrente di Bella Comba. L'accumulo ricopre invece il piede del cono di deiezione del torrente che proviene dal Monte Gruetta, con una evidente cresta allungata trasversalmente a esso.

I pochi affioramenti osservabili mostrano come l'accumulo consista essenzialmente di blocchi e grossi massi, con una matrice di sabbia e limo.

L'accumulo sin qui descritto è stato interpretato da quanti hanno studiato recentemente questo territorio come una morena deposta da un ghiacciaio vallivo formato dalla confluenza degli attuali ghiacciai del Triolet e di Pré de Bar. Non vi è però accordo quanto all'età della morena, che nei diversi autori varia dall'ultima espansione del Tardoglaciale (circa 10 000-11 000 anni fa) all'Olocene superiore (Subatlantico antico, circa 1000-3000 anni fa).

L'accumulo detritico grossolano dell'alta Val Ferret è in gran parte ricoperto da una vegetazione boscosa (larici, abeti rossi) ben sviluppata, spesso di apparente età ultracentenaria. Ciò ci consente di utilizzare le tecniche della dendrocronologia per tentare una sua datazione. La dendrocronologia è lo studio sistematico degli anelli di accrescimento degli alberi, allo scopo di ottenere datazioni assolute di reperti o di eventi di età storica e preistorica. Mediante tecniche di correlazio-

L'alta Val Ferret e il ghiacciaio del Triolet. È indicata la probabile area di origine della frana del 1717 presso l'Aiguille de l'Eboulement e il suo possibile percorso nel tratto più elevato del bacino del Triolet. La linea in rosso indica l'estensione massima raggiunta dal corpo di frana, la linea dentellata in blu l'estensione raggiunta dal ghiacciaio durante la fase di massima espansione all'inizio del secolo XIX. Probabilmente all'epoca della frana il ghiacciaio del Triolet aveva uno sviluppo non molto dissimile da quello massimo. Il dislivello e la distanza massimi percorsi dalla frana furono rispettivamente di 1860 metri e di 7 chilometri, la velocità raggiunta di oltre cento chilometri all'ora e il volume di 16-20 milioni di metri cubi. Si noti il percorso serpeggiante seguito dall'accumulo di frana che oscillò da un versante all'altro, risalì il versante meridionale della Val Ferret per almeno 60 metri nei pressi di Sagioàn, quindi deviò a destra, arrestandosi nella piana di Ferraché.

ne, utilizzando alberi viventi e resti lignei morti è stato possibile, in talune aree, costruire scale cronologiche estendentisi fino a oltre 8000 anni fa. Quando è utilizzata, come in questo caso, nello studio di problemi geomorfologici questa disciplina prende il nome di dendrogeomorfologia, mentre quando è impiegata per lo studio delle condizioni climatiche del passato è indicata come dendroclimatologia.

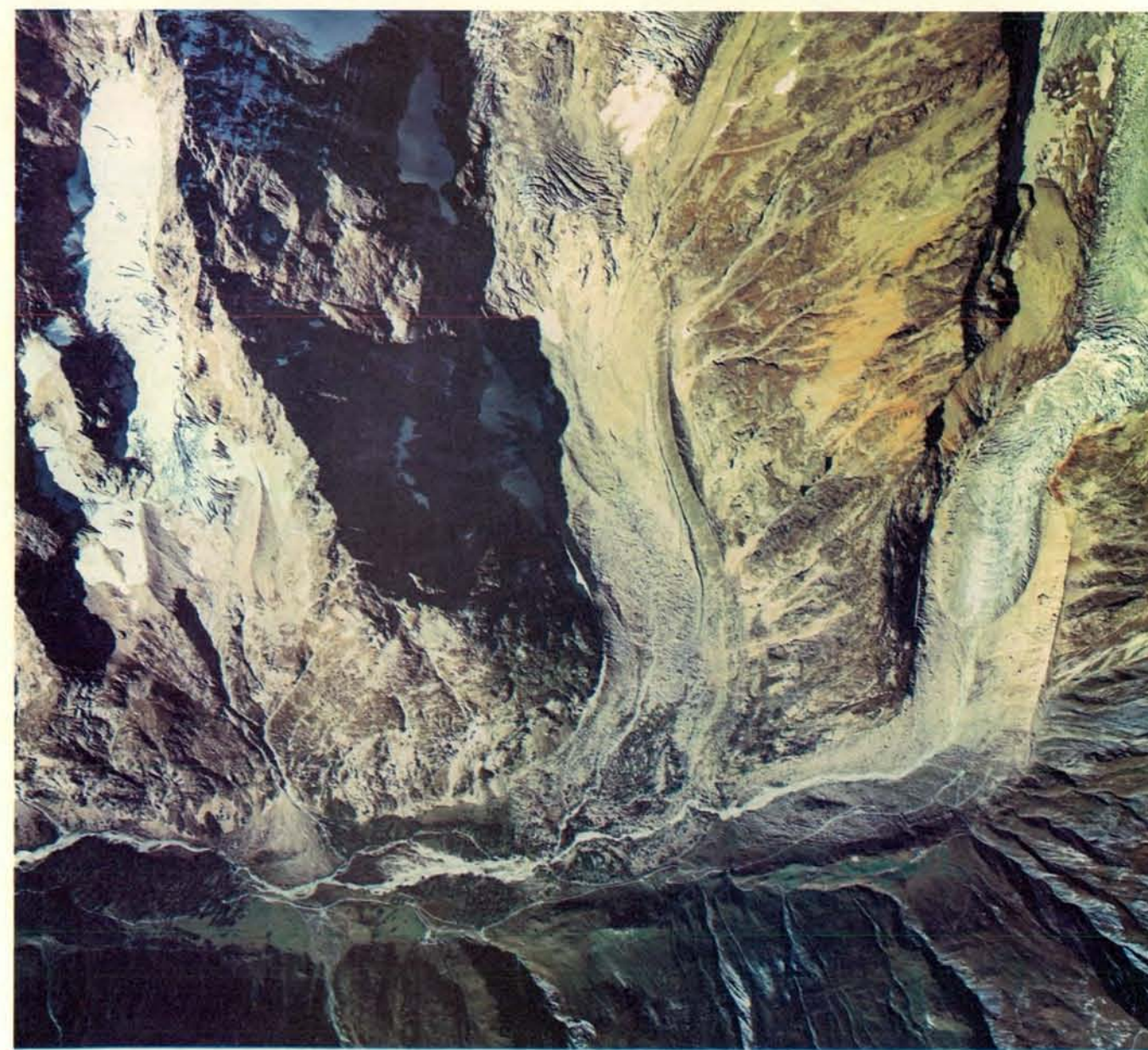
La dendrocronologia si basa sull'accrescimento discontinuo, con ritmo stagionale, degli alberi. In una sezione trasversale di un tronco si osservano anelli concentrici annuali, ciascuno dei quali è costituito da una parte interna di colore più chiaro, caratterizzata da cellule grandi a parete sottile (legno iniziale o primaverile) e da una parte esterna più scura e più densa, formata da cellule più piccole e a parete più spessa (legno tardivo o estivo). Il passaggio dal legno iniziale a quello tardivo in uno stesso anello è graduale, mentre quello dal legno tardivo di un anello al legno iniziale dell'anello successivo è estremamente brusco. Ciò consente una facile individuazione dei singoli anelli e una precisa misura del loro spessore. Salvo casi eccezionali (possibile mancanza di anelli annuali, possibile presenza di falsi anelli annuali) il semplice conteggio degli anelli in una sezione trasversale di un tronco o in una «carota» radiale estratta dal tronco, ci consente di conoscere l'età della parte di tronco considerata. Per determinare l'età di un albero occorre conteggiare gli anelli quanto più possibile alla base del tronco. L'albero più vecchio di un'associazione arborea ci permette di datare l'età minima di insediamento della stessa e quindi l'età minima di stabilizzazione della superficie topografica su cui si è sviluppata. Per approssimarci maggiormente all'età vera di stabilizzazione occorre aggiungere all'età dell'albero più vecchio l'intervallo di tempo necessario per l'insediamento e la germinazione degli alberi sulla nuova superficie topografica. Tale intervallo è variabile da specie a specie e da luogo a luogo, in funzione delle condizioni climatiche, edafiche, morfologiche, ecc.

In taluni casi un evento morfologico può essere direttamente datato con l'approssimazione di un anno, quando causa conseguenze dirette sulla modalità di accrescimento di un albero precedente all'evento e a esso sopravvissuto (frane, valanghe, alluvioni, ecc.), come si è visto nel caso della frana della Brenva illustrato nella figura a pagina 74.

Sull'accumulo detritico grossolano dell'alta Val Ferret l'albero più vecchio rinvenuto è stato un larice che possedeva almeno 232 anelli a una altezza di circa 1 metro dalla base. Il conteggio degli anelli è stato effettuato al microscopio binoculare su sottili carote estratte con il carotiere svedese Djos. Altri tronchi tagliati, presenti nelle vicinanze, mostravano oltre 200 anelli. Tutto ciò ci dice che sulla superficie dell'accumulo erano presenti alberi intorno alla metà del secolo XVIII. Se teniamo presente il tempo di insediamento e di sviluppo sino a una altezza di 50 centimetri osservato per i larici sviluppatisi sulla frana della Brenva (circa 15-20 anni), si può ritenere che la superficie dell'accumulo è stabile almeno dal 1727-1732 (o forse da qualche anno prima perché il dato di 232 anelli è stato ottenuto a 1 metro dal suolo). Poiché nelle valli di Courmayeur sono presenti larici vecchi anche 400-500 anni, si può ritenere che l'età ottenuta sia un'età minima ragionevolmente prossima a quella effettiva di stabilizzazione della superficie di accumulo, sempreché il bosco di larici e abeti rossi su di esso presente non sia stato nel passato interamente abbattuto dall'uomo o distrutto dalle valanghe e dagli incendi.

La dendrocronologia in questo caso lascia quindi adito a qualche dubbio in quanto non si ha la sicurezza di aver reperito l'albero più vecchio, l'intervallo di insediamento non è noto con precisione e non vi sono garanzie che l'associazione studiata sia la prima impiantatasi sopra la superficie dell'accumulo.

Un'altra possibilità di datazione dell'accumulo di blocchi dell'alta Val Ferret ci è offerta dalla lichenometria. Questa tecnica di datazione è stata introdotta da R. E. Beschel agli inizi degli anni cinquanta e dopo di allora è stata usata con successo, particolarmente per datare morene di età storica e preistorica, anche sino a circa 2000 anni.

I licheni crostosi sono forme vegetali estremamente diffuse sulle superfici rocciose, che colonizzano non appena vengono esposte. Le dimensioni dei licheni e il grado di copertura aumentano in funzione del tempo trascorso da quando una nuova superficie rocciosa è divenuta esposta e stabile. I licheni quindi possono essere usati per datare l'età delle superfici rocciose su cui crescono, quando ne sia nota la velocità di accrescimento. Questa varia da specie a specie e, per una stessa specie, da luogo a luogo, in funzione delle condizioni climatiche, dell'altitudine, dell'esposizione, del tipo di substrato roccioso e così via. La velocità di accrescimento varia inoltre nel tempo: dopo un interval-

Questa fotografia aerea, eseguita nella stessa occasione di quella a pagina 73, mostra l'alta Val Ferret, i ghiacciai del Triolet e di Pré de Bar, i loro apparati morenici recenti della «piccola età glaciale» (in questo caso corrispondente ai secoli XVIII e XIX) e l'accumulo di frana ricoperto da una folta vegetazione a bosco. Quest'ultimo è localmente interrotto da successivi depositi alluvionali.

lo necessario per l'insediamento e per raggiungere dimensioni macroscopiche (che può durare anche qualche decina d'anni) vi è una prima fase di rapido accrescimento seguita da una fase prolungata ad accrescimento più lento, gradualmente decrescente (in genere in modo esponenziale).

Operando in un territorio relativamente ristretto, entro limiti altimetrici definiti e su di un medesimo substrato roccioso, è possibile costruire una curva di accrescimento per una determinata specie. A questo scopo si misura il diametro minimo del lichene più grande a tallo circolare o subcircolare della specie prescelta, crescente sopra superfici di età nota, costituite dal medesimo tipo di roccia (per esempio pietre tombali, costruzioni datate, massi di frane, accumuli di piena alluvionale, morene e aree deglaciate prodottesi in età conosciuta) e si riportano i valori ottenuti in un diagramma dimensione/età. Si ottiene così una curva nella quale si distingue un tratto iniziale di rapido accrescimento e un tratto successivo ad accrescimento lento e decrescente. Misurando ora il diametro minimo del tallo più grande a contorno circolare della specie prescelta sopra superfici del medesimo substrato, ma di età sconosciuta, è possibile, per interpolazione o estrapolazione, usando la curva costruita, ottenere un'età minima di stabilizzazione delle stesse.

La considerazione fondamentale è che soltanto l'individuo più grande, e quindi ad accrescimento ottimale, potrà fornire un'età minima di stabilizzazione del substrato. I talli di dimensioni minori apparterranno o a successivi colonizzatori o a individui ad accrescimento più lento. Particolare attenzione deve quindi essere posta nella ricerca del tallo a dimensioni maggiori; i licheni di forma non circolare non debbono essere presi in considerazione, in particolare se a contorno lobato o comunque composti da due o più talli coalescenti.

Bisogna inoltre tenere presente che i diametri misurati esprimono un'età minima di stabilizzazione della superficie rocciosa e che tale età si approssima maggiormente all'età vera quanto più estesa è la superficie presa in considerazione, perché in tal caso è maggiore il numero di licheni misurati e aumentano le probabilità di rinvenire il lichene più grande e quindi più antico possibile. Quando si opera, invece, su pietre tombali o altre superfici molto limitate si ottengono valori sicuramente inferiori rispetto a quelli che si otterrebbero analizzando superfici più estese, quali morene, superfici di frana, ecc.

La precisione del metodo dipende dalla quantità dei punti di controllo e dalla velocità di accrescimento e decresce comunque con l'aumentare dell'età. Per gli ultimi 3-4 secoli si possono ottenere date con un'approssimazione dell'ordine di ± 5 anni.

Nel corso di studi sulle variazioni delle fronti dei ghiacciai del versante italiano del Bianco è stata costruita una curva di accrescimento per il lichene *Rhizocarpon geographicum* su substrato di granito del Monte Bianco. Si tratta del comune lichene di colore giallo verdastro con aspetto reticolato e contorno di colore nero, presente su pressoché tutti i massi granitici del territorio e che raggiunge, nell'area studiata, diametri fino a oltre 250 millimetri. Per la costruzione della curva di accrescimento come superfici di età nota sono state prese in considerazione le lastre tombali del cimitero di Courmayeur, costruzioni e manufatti di età conosciuta, i massi della frana della Brenva del 1920 nonché morene di età storica (risalenti ai secoli XIX e XX) dei ghiacciai del Triolet e di Pré de Bar. La velocità di accrescimento riscontrata, superata la fase iniziale, è di circa 0,2 millimetri all'anno.

Lo studio della copertura lichenica sull'accumulo dell'alta Val Ferret ha rivelato come valore massimo per *R. geographicum* 75 millimetri; questo valore, riportato sull'estrapolazione della curva costruita, indica un'età riferibile al decennio 1717-1727. I licheni confermano quindi per l'accumulo di massi dell'alta Val Ferret un'età che risale all'inizio del XVIII secolo.

A questo punto si hanno sufficienti elementi sia morfologici che cronologici per affermare che l'accumulo di massi dell'alta Val Ferret non è una morena deposta da un ghiacciaio, ma un accumulo di frana, la cui età, ubicazione ed estensione coincidono e con quelle riferite dai documenti storici per la frana del Triolet.

Infatti la superficie dell'accumulo dell'alta Val Ferret presenta un gradiente longitudinale estremamente basso (20 per mille), di un ordine di grandezza inferiore a quello presentato dalle morene deposte dai vicini ghiacciai del Monte Bianco. I rilievi longitudinali osservati non hanno continuità né mostrano una distribuzione simmetrica sui due lati della valle, quale di norma possiedono le morene laterali. L'accumulo di blocchi ha in pianta un andamento lievemente serpeggiante e risale leggermente ora su un lato ora su quello opposto della valle. La superficie ha un aspetto caotico con depressioni e rilievi irregolarmente disposti, mentre la granulometria è eccezionalmente grossolana.

Sia gli alberi che i licheni indicano un'età riferibile all'inizio del secolo XVIII. È pertanto da scartarsi un'attribuzione dell'accumulo a un'espansione glaciale antica, tardoglaciale o, comunque, preistorica. Sebbene l'inizio del secolo XVIII sia stato un periodo di espansione glaciale (nell'ambito della «piccola età glaciale», che si verificò tra il XVI e il XIX secolo) da quanto risulta dagli studi da noi fatti su gli altri ghiacciai italiani del Monte Bianco, l'accumulo di blocchi dell'alta Val Ferret si estende a una distanza insolitamente grande rispetto alle morene deposte all'inizio del secolo XIX. Queste ultime sono in generale le più esterne, testimoniano cioè la massima avanzata glaciale verificatasi durante l'Olocene. Quando si sono osservate morene più antiche, del XVIII secolo o precedenti, esse giacciono sempre immediatamente all'esterno di quelle del secolo scorso.

La fronte dell'accumulo di frana del Triolet vista da valle, nella piana di Ferraché. Il margine dell'accumulo si erge netto di qualche metro sulla piana circostante, simulando un arco morenico. Esso fu infatti scambiato in precedenza per una morena deposta dai ghiacciai del Triolet e di Pré de Bar congiunti, in una fase preistorica di maggiore estensione.

Si può quindi concludere che l'accumulo dell'alta Val Ferret è in realtà l'accumulo della frana catastrofica del 1717, descritta dal Pennard e dal De Tillier. Ciò ci consente di definire questo evento nelle sue modalità con maggiore precisione. L'accumulo di frana si può stimare ammonti almeno a 16-20 milioni di metri cubi, attribuendo al corpo di frana una superficie di 2,7 chilometri quadrati e uno spessore medio di circa 6-8 metri. Se, come vuole una tradizione che risale almeno ai tempi di Forbes (1843), l'area di origine della frana è stata l'Aiguille de l'Eboulement, la distanza massima percorsa è stata di oltre 7 chilometri, il dislivello massimo di 1860 metri, la velocità raggiunta, a giudicare dal dislivello minimo risalito sul versante opposto (circa 60 metri nei pressi di Sagioàn), di oltre 100 chilometri all'ora. È verosimile che all'atto della frana, come si è già accennato, il ghiacciaio del Triolet si trovasse in una fase di maggiore espansione rispetto a quella attuale, raggiungendo probabilmente la valle principale.

Per dimensioni volumetriche e soprattutto per la distanza percorsa, la frana del Triolet si colloca tra le maggiori avvenute nelle Alpi.

Percorrendo le due valli italiane del Monte Bianco e studiando le oscillazioni glaciali, ci si è resi gradualmente conto che le frane della Brenva e del Triolet non sono state le sole in tempi olocenici in queste valli. La superficie del ghiacciaio del Miage è interamente ricoperta da detriti molto grossolani; sulle foto aeree si possono chiaramente riconoscere distinti accumuli di frana, per il colore diverso delle rocce di cui sono composti e per la forma a lobo allungato, dovuta in parte alla deformazione indotta dal flusso del ghiacciaio. Anche le carte topografiche antiche mostrano spesso questa caratteristica copertura per frana del ghiacciaio del Miage.

Modeste frane di crollo (nicchie di distacco e sottostanti accumuli di frana) sono visibili in più punti al piede del versante meridionale del Monte Bianco, come per esempio in Val Veni tra la fronte del Miage e le case del Peutérey o in Val Ferret presso Frébouze.

In Val Ferret numerosi accumuli di blocchi sparsi a semicerchio sul fondovalle ai piedi del versante meridionale del gruppo del Bianco, quali quello di fronte al ghiacciaio di Frébouze e che risale sul versante opposto subito a monte di La Vachey, e quelli che si susseguono quasi con continuità da Planpincieux a La Vachey sono riconducibili ad antiche frane. La copertura lichenica indica per essi un'età variabile da 150 a oltre 600 anni fa.

Al margine dell'accumulo della frana della Brenva del 1920 sono visibili altri massi più antichi che fanno pensare che questo fenomeno si sia qui ripetuto altre volte nel passato. Infine la zona intorno alla stazione della funivia del Pavillon è disseminata di grandi massi granitici non associati a depositi glaciali, anch'essi di probabile origine franosa, sebbene di età molto antica, a giudicare dalle dimensioni dei licheni, tra le maggiori osservate. Si può quindi concludere che nel versante meridionale del Monte Bianco, particolarmente nella bassa Val Veni e in Val Ferret, le frane di grandi dimensioni siano un fenomeno relativamente comune e ripetitivo, con un tempo di ritorno dell'ordine del secolo.

Come è stato detto all'inizio, le frane di dimensioni catastrofiche sono un fenomeno proprio, in generale, dell'area alpina e l'esempio portato delle valli di Courmayeur vuole quindi avere il significato di un più generale richiamo al potenziale rischio di frana presente nel territorio alpino, in particolare nelle aree più elevate. Si è voluto inoltre indicare una possibile modalità di studio per ubicare con maggiore precisione e datare le frane verificatesi nel recente passato geologico, al fine di dare una prima indicazione di massima delle aree potenzialmente più esposte a tale rischio e concentrare conseguentemente su di esse gli studi geologico-tecnici necessari. In particolare, nel caso di piani di progettazione territoriale e di sviluppo delle valli alpine, lo studio geologico delle vicende verificatesi durante l'Olocene può indicare quali siano le aree che più frequentemente sono state esposte a frane di tipo catastrofico e quindi orientare opportunamente le scelte, così da evitare che tali aree divengano sedi di insediamenti permanenti o comunque siano destinate ad attività che possano determinare un'elevata concentrazione della popolazione.

Curva di accrescimento del lichene *Rhizocarpon geographicum* costruita su misure effettuate nelle valli di Courmayeur. In ordinate sono riportati i diametri minimi dei talli circolari o subcircolari più grandi osservati su superfici di età nota, quali pietre tombali, costruzioni datate, morene o superfici di frana di età conosciuta e in ascisse le età in anni. I valori ottenuti, interpretabili come valori minimi, consentono di tracciare una curva che può essere usata per interpolazione o per estrapolazione per datare superfici di età ignota. Questa tecnica di datazione, detta «lichenometria», è stata introdotta agli inizi degli anni cinquanta da Beschel e successivamente usata con buoni risultati, particolarmente per datare morene.

La formazione della Terra da planetesimali

In base a una simulazione al calcolatore, i pianeti interni del sistema solare potrebbero essersi formati per aggregazione da uno sciame di piccoli corpi che collidevano orbitando attorno al Sole

di George W. Wetherill

L'accrescimento dei pianeti interni del sistema solare dalla collisione di corpi più piccoli chiamati planetesimali è stato simulato al calcolatore da Larry P. Cox del Massachusetts Institute of Technology. La simulazione ha inizio con 100 corpi di uguale massa. Secondo le leggi di Keplero, ciascun corpo segue un'orbita ellittica attorno al Sole. Ogni orbita è stata specificata da due parametri: il semiasse maggiore *(cartina in alto)* e l'eccentricità *(cartina in basso)*. Il semiasse maggiore misura la distanza dell'orbita dal Sole in unità astronomiche, U. A. (Una U.A. è la distanza media fra Terra e Sole, pari a circa $1,5 \times 10^8$ chilometri.) L'eccentricità misura quanto un ellisse si discosta dalla forma circolare. (L'eccentricità uguale a 0 rappresenta un cerchio; l'eccentricità uguale a 1 rappresenta una parabola in cui il semiasse maggiore si protrae all'infinito.) Sebbene inizialmente i parametri delle orbite venissero generati a caso, la massima eccentricità iniziale è stata scelta pari a 0,15 e i semiassi maggiori tra valori compresi fra 0,5 e 1,5 U.A. La simulazione è bidimensionale, ossia pone tutte le orbite su un piano. Le collisioni fra i corpi nel piano sono più probabili che non in uno spazio tridimensionale; di conseguenza l'accrescimento in questa simulazione «piatta» è completo in circa soli 61 000 anni. I quadratini nelle cartine indicano le collisioni da cui deriva un'aggregazione. Le discontinuità in alcuni quadratini nella cartina inferiore indicano forti diminuzioni dell'eccentricità. Alla fine rimangono sei pianeti. I diversi colori contraddistinguono le differenti sequenze secondo cui essi si accrescono. Il pianeta che rimane più vicino al Sole è un planetesimale che non ha mai subito collisioni, sebbene le divagazioni del suo moto mostrino che la sua orbita è stata ripetutamente perturbata da collisioni mancate con altri corpi. Gli altri cinque pianeti sono di dimensioni maggiori e potrebbero essere chiamati con un po' di fantasia Mercurio, Venere, Terra, Luna e Marte. Il successo della simulazione effettuata da Cox sta nel fatto che da essa emergono corpi che assomigliano alla Terra per massa e tipo di orbita. L'asse del tempo è stato distorto in modo che un numero uguale di eventi perturbanti (collisioni avvenute e mancate) occupi uguali intervalli di questi stampati.

Per raggiungere una comprensione totale sulle modalità di formazione della Terra è necessario capire come il Sole si sia condensato da una nube di gas e polvere e come una piccola parte di questa materia sia riuscita a sfuggire al Sole e a formare così i pianeti. Queste conoscenze non saranno facili da raggiungere. Da un lato le simulazioni al calcolatore suggeriscono che se una nube di questo tipo, ossia una nebulosa solare, è abbastanza densa da subire un collasso gravitazionale, darà origine a numerosi sistemi stellari multipli invece che a una singola stella. Inoltre anche se si fosse capito tutto sulle modalità di condensazione del Sole, questo potrebbe non chiarire completamente l'origine dei pianeti interni (ossia i pianeti di tipo terrestre, Mercurio, Venere, Terra e Marte). La massa dei pianeti interni è solo lo 0,0005 per cento della massa del Sole. Ovviamente non è ragionevole aspettarsi di raggiungere una conoscenza dettagliata di un processo così preciso e complesso come la formazione della Terra, attraverso lo studio delle sole proprietà generali della nebulosa solare.

Sembra più promettente un altro metodo. A parte la complessità implicita dell'origine del sistema solare da una nebulosa solare, esistono solo due altri modi fondamentali in cui possono essersi formati i pianeti interni. Possono essere derivati da una concentrazione locale di gas e polvere della nebulosa condensatisi in un pianeta per la loro stessa forza gravitazionale, oppure la nube può essersi formata per l'accumulo di piccoli corpi (planetesimali), formatisi in precedenza per condensazione da gas e polvere cosmica. Creando simulazioni al calcolatore di queste ipotesi in alternativa, i teorici stanno migliorando la conoscenza sulle modalità di formazione del sistema solare. Per esempio, sta diventando possibile specificare le condizioni che hanno portato alla formazione dei pianeti nella nebulosa solare e distinguerle dalle condizioni sfavorevoli a questo sviluppo. Sta inoltre diventando possibile determinare che cosa ciascuna ipotesi è in grado di prevedere riguardo a proprietà come la temperatura iniziale della Terra. Si può essere ottimisti sul fatto che il confronto tra le previsioni e i dati geologici renderà evidente quale alternativa rappresenti la via più probabile della formazione dei pianeti interni.

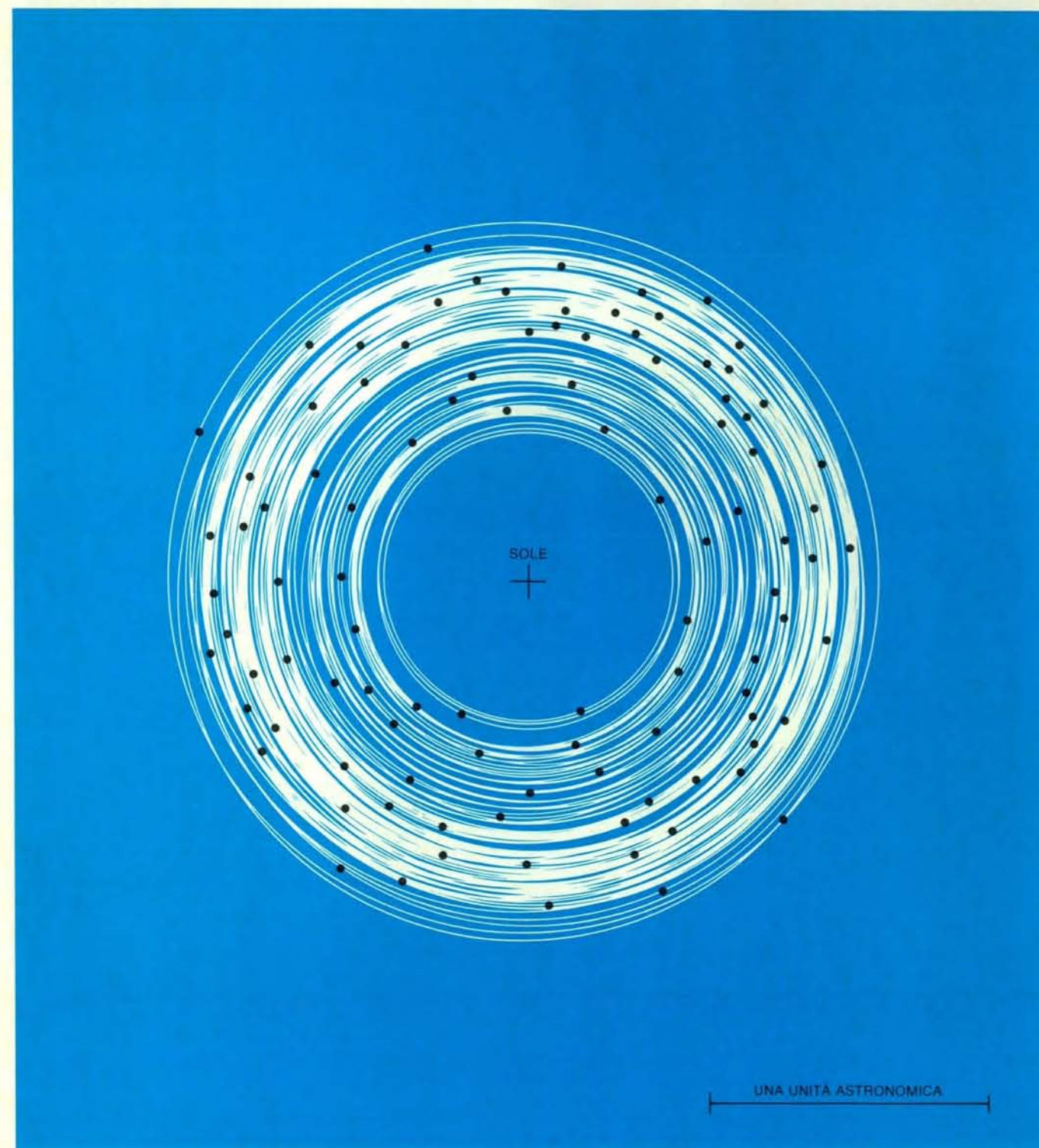
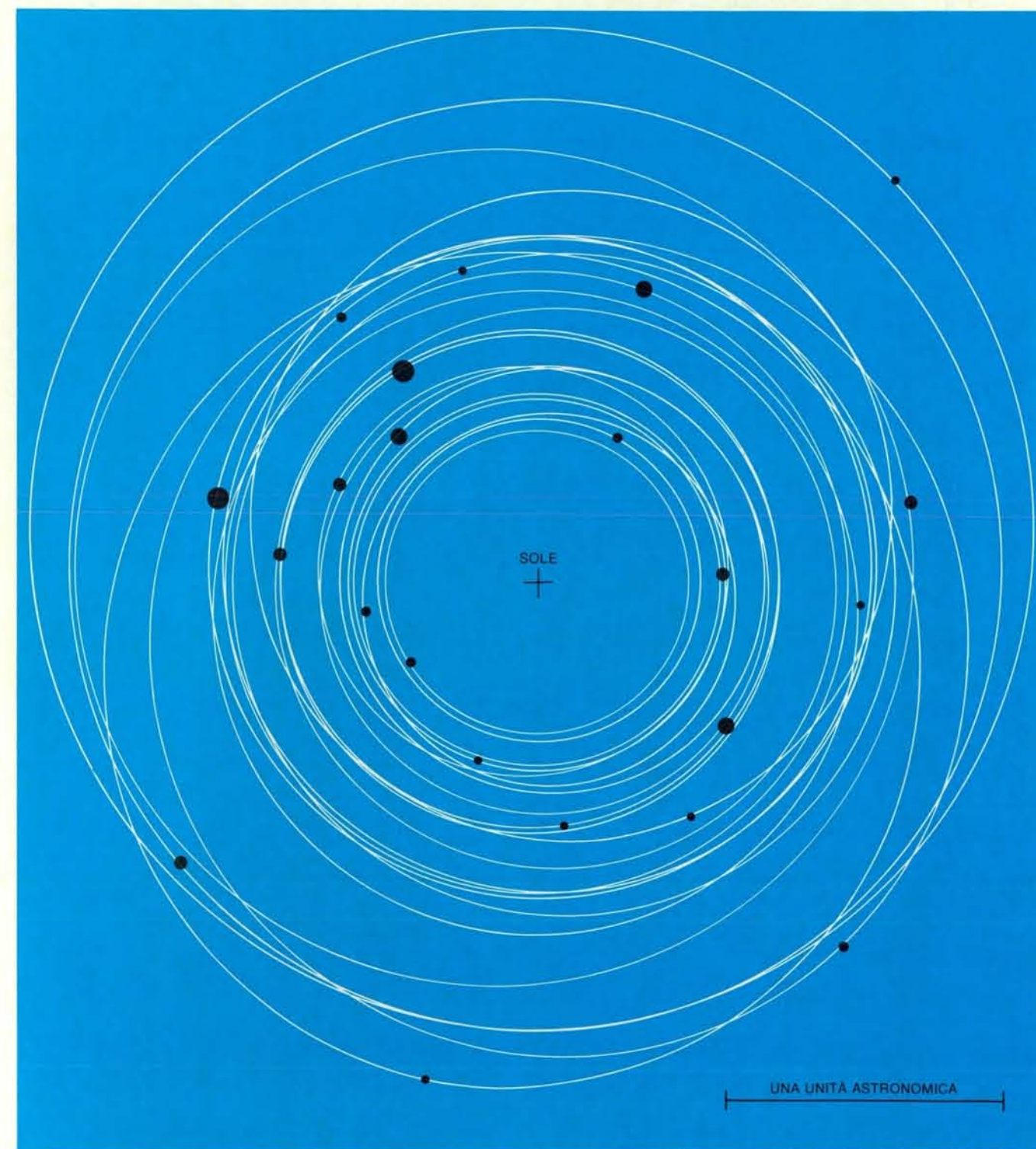
In questo articolo esporrò le attuali conoscenze relative alla formazione dei pianeti per accumulo di planetesimali. È troppo presto per decidere se la Terra si sia formata in questo modo o per collasso gravitazionale. Appare abbastanza probabile, tuttavia, che i pianeti interni del sistema solare si siano formati da planetesimali. Se le cose stavano così, le simulazioni indicano che il processo di formazione era pressoché completo dopo circa 100 milioni di anni dal suo inizio e che la Terra appena formatasi aveva una temperatura così elevata che gran parte del suo interno era allo stato fuso.

Il primo problema da considerare nella costruzione di pianeti da planetesimali consiste nel verificare se quando due corpi collidono tendono ad aggregarsi formando un corpo di maggiori dimensioni. È ovvio che non sempre una collisione si conclude con un fenomeno di aggregazione. Per esempio se due corpi rocciosi di un metro di diametro collidono alla velocità di 1000 chilometri all'ora (0,3 chilometri al secondo) si frantumano e si allontanano reciprocamente. Questo non è sicuramente un sistema che si possa ipotizzare per la formazione dei pianeti di tipo terrestre.

Una collisione a velocità relativamente bassa consente invece ai corpi di aggregarsi. Quanto deve essere bassa questa velocità? Nel rispondere a questa domanda diventa importante la velocità di fuga che si oppone all'attrazione gravita-

zionale, ossia la velocità relativa minima necessaria perché due corpi non risentano della reciproca attrazione gravitazionale. Per esempio per sfuggire dalla Terra un proiettile sparato dalla superficie deve avere una velocità di circa 11 chilometri al secondo; per sfuggire al campo gravitazionale della Luna è sufficiente invece una velocità di 2,4 chilometri al secondo. Se due planetesimali collidessero alla loro esatta velocità di fuga e se non vi fosse dispersione di energia, essi avrebbero forza sufficiente per rimbalzare e allontanarsi su traiettorie indipendenti.

In una collisione reale i planetesimali perderebbero energia perché due corpi che collidono si deformano e si rompono. La perdita di energia avverrebbe sotto forma di dispersione di calore. I due planetesimali che collidessero esattamente alla loro velocità di fuga, non sarebbero pertanto in grado di sfuggirsi reciprocamente e i loro frammenti si unirebbero a formare un corpo di dimensioni maggiori. E infatti gli esperimenti fatti da William K. Hartmann del Planetary Science Institute di Tucson (Arizona) e i calcoli numerici eseguiti da Thomas J. Ahrens e John D. O'Keefe del California Institute of Technology hanno dimostrato che quando la velocità di impatto è due o, in alcuni casi, anche tre volte la velocità di fuga, i planetesimali possono aggregarsi.

Studi successivi ampliano questa linea di ragionamento. La velocità di impatto di due corpi è determinata dalla velocità relativa dei corpi quando distavano l'uno dall'altro, a cui va aggiunta la velocità che acquisiscono per l'accelerazione dovuta all'attrazione di gravità. Se, quando i due corpi entrano in collisione, la velocità relativa è due volte la velocità di fuga, la velocità relativa quando i due corpi erano a grande distanza era uguale a $\sqrt{3}$ (circa 1,7) volte la velocità di fuga. Se la velocità al momento della collisione è tre volte la velocità di fuga, la velocità

Anche la simulazione tridimensionale relativa alla formazione del sistema solare interno è cominciata con 100 planetesimali. A ciascuno di essi è stata attribuita una massa di 1,2 × 10^{26} grammi, in modo che la massa totale (1,2 × 10^{28} grammi) fosse uguale a quella dei pianeti interni da Mercurio a Marte, compresi i loro satelliti. Anche in questo caso i parametri iniziali delle orbite sono stati generati in modo casuale. L'illustrazione mostra le ellissi iniziali, ma non indica la loro inclinazione reciproca. Poiché si pensa che i planetesimali si siano formati in un sottile strato di polvere cosmica sul piano centrale della nube da cui si è condensato il sistema solare, le inclinazioni sono inferiori ai cinque gradi.

Lo stadio intermedio della simulazione mostra lo sviluppo del sistema solare interno dopo 30,2 milioni di anni, allorché rimangono 22 corpi. Le loro orbite sono diventate più marcatamente ellittiche e le velocità reciproche sono aumentate. Inoltre si è ampliata la gamma delle distanze dal Sole. Le orbite più interne sono più vicine al Sole, mentre le più esterne sono più lontane. L'inclinazione reciproca tra le orbite ha notevolmente diminuito le probabilità di collisione rispetto alle collisioni mancate. Pertanto devono passare 10 milioni di anni, in questa simulazione tridimensionale, prima che il processo giunga alla fine. Nella simulazione bidimensionale basterebbero invece 61 000 anni.

relativa, quando erano distanti, era $2\sqrt{2}$ (circa 2,8) volte la velocità di fuga. Pertanto la regola che due corpi possono aggregarsi solo se la loro velocità relativa al momento dell'impatto è inferiore di due o tre volte la velocità di fuga si applica anche alla velocità relativa di due corpi che si muovono indipendentemente molto prima della loro collisione.

In alcune simulazioni di formazione dei pianeti da planetesimali, la velocità relativa di due planetesimali qualsiasi è stata considerata un parametro indipendente: si è assunto che la velocità possa avere sempre un valore abbastanza basso da consentire l'aggregazione dei planetesimali. D'altra parte la velocità può anche essere troppo bassa. Le leggi sul moto dei pianeti formulate per la prima volta da Keplero specificano che l'orbita di un corpo attorno al Sole ha la forma di un'ellisse. La velocità in ogni punto di questa orbita è determinata dalle dimensioni e dalla forma dell'ellisse. Ne consegue che la velocità relativa di due corpi in orbita attorno al Sole tenderà a essere elevata, a meno che le orbite non siano simili in dimensione, forma e orientazione. Ne risulta che la scelta di una velocità relativa bassa limita le possibilità di aggregazione e di accrescimento solo a una piccola parte dei planetesimali, cioè a quelli che hanno orbite simili. Il processo di formazione dei pianeti avrà fine quando questi planetesimali si saranno aggregati e il sistema solare interno simulato assomiglierà agli anelli di Saturno piuttosto che essere caratterizzato da un gruppo di quattro pianeti di tipo terrestre in uno spazio che si presenta per il resto vuoto.

Ovviamente la scelta di una velocità relativa vicina alla velocità di fuga sarebbe la strategia migliore per una simulazione riuscita. Consentirebbe infatti ai planetesimali in collisione di aggregarsi e nello stesso tempo renderebbe massimo il numero di potenziali corpi in grado di collidere. Si pone, però, un'altro problema. Quando i corpi aumentano di dimensioni cresce anche la velocità di fuga, che in un corpo di densità uniforme è proporzionale al raggio. Se un planetesimale con un raggio di un metro avesse una velocità di fuga di 0,15 centimetri al secondo, un planetesimale con un raggio di un chilometro avrebbe una velocità di fuga di 1,5 metri al secondo, e un planetesimale con un raggio di 1000 chilometri avrebbe una velocità di fuga di 1,5 chilometri al secondo. Ciascuno di questi corpi può continuare a crescere per aggregazione se collide con un corpo che abbia una velocità non superiore al doppio o al triplo della velocità di fuga. Pertanto la velocità relativa dei planetesimali deve andare di pari passo con la loro crescita e con la velocità di fuga in aumento, per rimanere entro la gamma dei valori ottimali.

La possibilità che la velocità dei planetesimali possa tenere naturalmente il passo con la crescente velocità di fuga fu avanzata nel 1950 da L. E. Gurevich e A. I. Lebedinskii dell'Istituto di geofisica applicata di Mosca. L'idea fu poi ampiamente sviluppata da Victor S. Safronov dello stesso istituto durante gli anni sessanta. Il lavoro di Safronov fu analitico: egli sviluppò le espressioni matematiche dell'evoluzione di entità quale la velocità relativa media dei membri di uno sciame di planetesimali. Il suo modello più semplice descriveva lo sciame di planetesimali in assenza di ogni effetto dovuto ai gas della nebulosa solare.

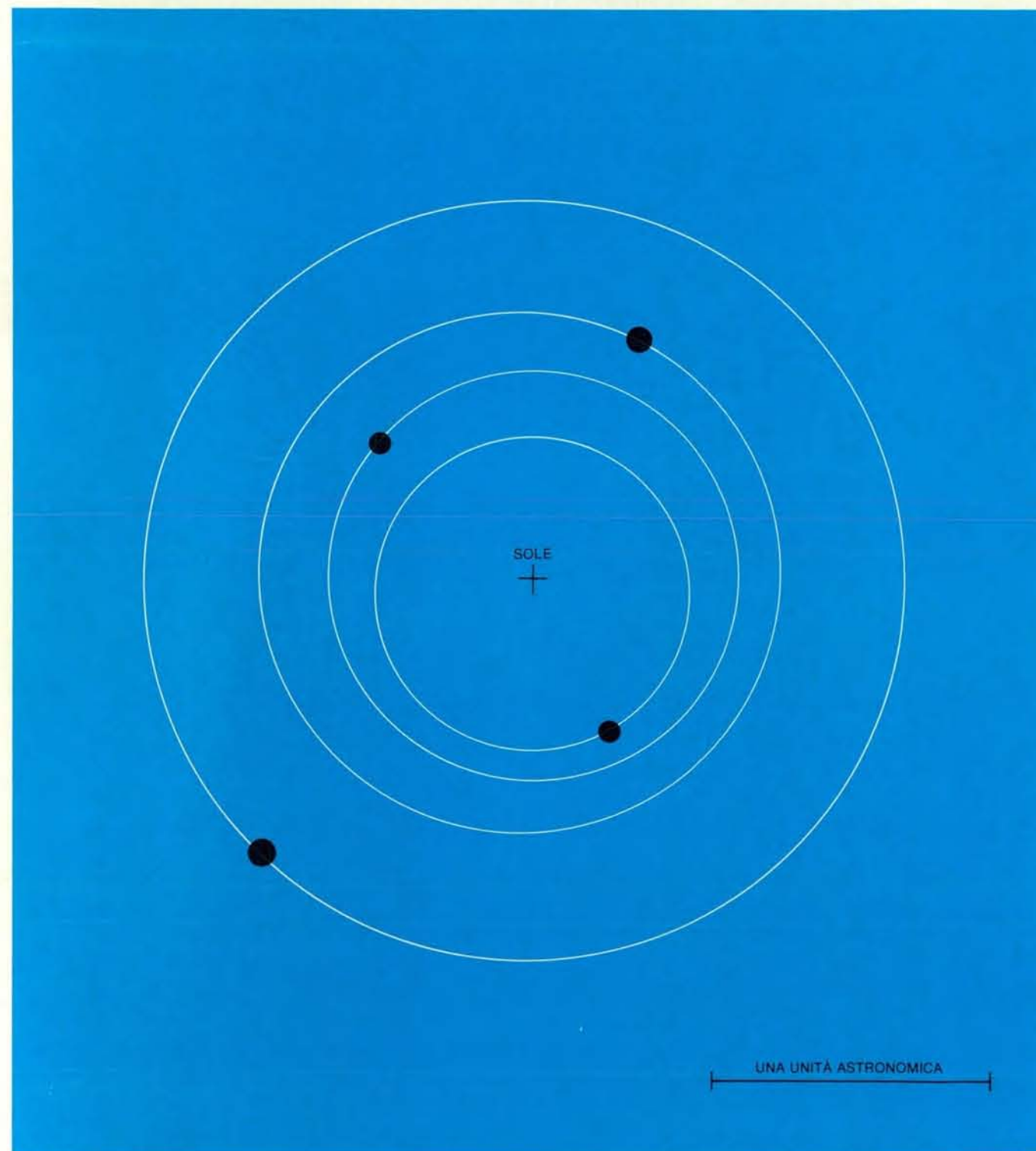
Le velocità relative dei corpi dello sciame è determinata dall'equilibrio fra due effetti opposti. Quando i corpi passano l'uno vicino all'altro senza collidere, l'attrazione gravitazionale perturba le loro orbite. Questa perturbazione modifica la velocità relativa alla quale ciascun corpo successivamente incontra altri corpi che si muovono in orbite intersecantisi. Talvolta la velocità relativa dopo che i corpi si sono sfiorati è maggiore di quanto altrimenti sarebbe stata; talvolta invece è minore. Il risultato netto, tuttavia, è un aumento del valore medio della velocità relativa. Le collisioni hanno l'effetto opposto. Tendono a rendere le orbite dei planetesimali più circolari e quindi più simili; ne consegue che le velocità relative tendono a diminuire. In un qualsiasi momento dato, pertanto, l'aumento della velocità determinato dalle collisioni mancate tende a essere controbilanciato dalla diminuzione di velocità dovuta alle collisioni.

Ciascuno di questi effetti può essere considerato in maggior dettaglio, e in questo ci aiutano le leggi di Newton sul moto insieme con la teoria dei moti casuali che descrive la natura essenzialmente casuale delle interazioni fra planetesimali. Nel caso di piccole perturbazioni dovute a collisioni mancate, il cambiamento di velocità di un corpo di piccole dimensioni (un «proiettile»), che incontra un corpo di dimensioni maggiori (un «bersaglio»), risulta dipendere da $\sqrt{M/D}$, in cui M è la massa del bersaglio e D la distanza nel momento di massimo avvicinamento. Nei casi limite la distanza D può essere uguale a R, il raggio del bersaglio. Inoltre la massa M dipende da R^3. L'espressione $\sqrt{M/D}$ può quindi essere trasformata in $\sqrt{R^3/R}$ o semplicemente in R. In breve, quindi, il cambiamento di velocità del proiettile è direttamente proporzionale al raggio, proprio come la velocità di fuga. Questa relazione implica che l'aumento della velocità relativa che accompagna un certo numero di collisioni mancate procede di pari passo con l'aumento della velocità di fuga provocato dalla crescita dei corpi.

Anche «incontri» più distanti perturbano le orbite. In realtà, gli incontri sono significativi a qualsiasi distanza dal bersaglio inferiore al raggio di quella che si potrebbe chiamare sfera di influenza gravitazionale del bersaglio. All'interno di questo volume il moto del proiettile è dominato dal campo gravitazionale del bersaglio. Al di fuori di questo volume il moto è dominato dal campo gravitazionale del Sole. Se viene fatta la media delle variazioni di velocità causate da incontri distanti su tutte le possibili distanze di incontro entro il limite della sfera di influenza (circa 75 volte il raggio del bersaglio), si ottiene un risultato che è più o meno in proporzione fissa con le variazioni causate da incontri a distanze dell'ordine di un raggio del bersaglio. Di conseguenza anche gli aumenti di velocità provocati dagli incontri distanti crescono con la velocità di fuga.

L'aumento di velocità dovuto a incontri a qualsiasi distanza compresa fra un raggio e 75 raggi è più che sufficiente a controbilanciare le diminuzioni dovute alle collisioni. E infatti l'aumento dell'area della sezione trasversale di ciascun bersaglio, man mano che questo cresce, aumenta la frequenza delle collisioni. Nello stesso tempo, tuttavia, l'aumento di massa del bersaglio estende la sua sfera di influenza gravitazionale e di conseguenza aumenta la probabilità di un incontro perturbatore. Il risultato finale è notevole: la velocità relativa di uno sciame di planetesimali va effettivamente di pari passo con la velocità di fuga. Questo sembra essere proprio il risultato necessario affinché i planetesimali continuino ad aggregarsi all'aumentare delle dimensioni di ciascun corpuscolo. Inoltre, per valori plausibili di perdite di energia nelle collisioni, la velocità relativa è circa uguale alla velocità di fuga.

All'inizio del processo, secondo il modello di Safronov, i planetesimali sono di piccole dimensioni; ciascuno ha un raggio di pochi chilometri. Anche la velocità di fuga ha un valore basso e la velocità relativa ha un valore abbastanza vicino a quello della velocità di fuga. Solo corpi con orbite simili possono collidere e aggregarsi. Tuttavia questo fenomeno di aggregazione procede in tutta la regione del sistema solare che alla fine verrà occupata dai pianeti interni. La velocità di fuga cresce all'aumentare delle dimensioni dei corpi; tuttavia il processo di aggregazione continua in quanto la velocità relativa è prossima alla velocità di fuga. In accordo con le leggi di Keplero, l'aumento della velocità relativa è associato all'aumento dell'eccentricità delle orbite. In altre parole ciascuna orbita diventa un'ellisse sempre più pronunciata. Questa tendenza permette a ciascun planetesimale di raggiungere corpi sempre più lontani. Il numero di potenziali collisioni si mantiene perciò costante per ciascun planetesimale, anche se la quantità totale dei corpi diminuisce e i corpi diventano sempre più distanziati l'uno dall'altro.

Che cosa resta da capire prima di poter credere che la Terra e gli altri pianeti di tipo terrestre si siano davvero formati in questo modo? Ci si potrebbe chiedere come si siano formati i piccoli planetesimali che compaiono all'inizio nel modello di Safronov. Gli studiosi sono ormai d'accordo nel ritenere che questo non sia un problema serio. Stuart J. Weidenschilling del Planetary Science Institute ha dimostrato che le deboli attrazioni elettriche chiamate forze di van der Waals possono benissimo aver causato l'aggregazione dei granelli di polvere nella nebulosa solare, con formazione di corpi del diametro di un centimetro. Si assume che la nebulosa avesse un moto di rotazione. La resistenza esercitata su questi corpi da parte dei gas in rotazione ne avrebbe provocato l'addensamento in corrispondenza del piano centrale della nebulosa. (Il piano centrale è perpendicolare all'asse di rotazione della nebulosa.) Il risultato sarebbe stato la formazione di un sottile strato centrale di polvere. Alcuni corpi sarebbero cresciuti in questa zona per collisioni. Quando quelli di dimensioni maggiori avessero raggiunto un diametro di circa un metro, la concentrazione di materia nello strato di polvere avrebbe reso lo strato gravitazionalmente instabile. I corpi si sarebbero quindi aggregati alla polvere rimanente, formando planetesimali di circa un chilometro di diametro. Il processo è stato descritto da diversi studiosi negli ultimi 30 anni. Questi sono i plane-

Lo stadio finale della simulazione mostra quattro pianeti completamente formati, isolati in orbite quasi circolari attorno al Sole. Tutti i planetesimali si sono aggregati. Il pianeta di maggiori dimensioni, il quarto in ordine di distanza dal Sole, si è formato a partire da 34 corpi originari. Sebbene l'illustrazione mostri la simulazione dopo 441 milioni di anni, il processo è quasi ultimato molto prima. Dopo 79 milioni di anni ci sono 11 corpi e dopo 151 milioni di anni ce ne sono sei. Questo suggerisce che la Terra si sia accresciuta in 100 milioni di anni circa. La simulazione è stata effettuata dall'autore con un calcolatore del Department of Terrestrial Magnetism della Carnegie Institution di Washington.

tesimali su cui si basa la teoria di Safronov. Ne sono necessari circa 10^{13} per formare i pianeti interni.

Altri problemi sono un po' meno semplici. Per esempio, il modello più semplice di Safronov descrive uno sciame di planetesimali che all'inizio hanno tutti la stessa massa. Nel 1962 egli incluse nel suo modello il caso di corpi con massa diversa. Fu in grado di dimostrare che la velocità relativa si accorda con la velocità di fuga dei corpi di maggiori dimensioni solo nel caso in cui questi corpi costituiscano una percentuale elevata della massa del sistema. In questo caso i corpi più piccoli si frantumano nel corso delle collisioni, ma i più grandi si accrescono. Dato però che il sistema si evolve da un ampio numero di minuscoli planetesimali, è possibile che non si ottenga mai una concentrazione della massa in corpi di maggiori dimensioni. Per trattare a fondo questo problema bisognerebbe calcolare contemporaneamente l'evoluzione della velocità relativa e l'evoluzione della distribuzione di massa. La teoria di Safronov non è adeguata per questo tipo di calcoli.

Richard J. Greenberg e i suoi colleghi del Planetary Science Institute hanno eseguito calcoli importanti a questo riguardo. Il loro lavoro si è limitato ai primi stadi di crescita in uno sciame di planetesimali, cominciando dal momento in cui la maggior parte della massa dello sciame è concentrata in planetesimali di circa un chilometro di diametro. La velocità relativa di questi corpi è solo di pochi metri al secondo, valore paragonabile con la velocità di fuga. Secondo i calcoli di Greenberg alcuni dei corpi entrano in collisione e si aggregano e, dopo circa 15 000 anni, un numero ridotto di questi si è accresciuto fino a raggiungere i 100 chilometri di diametro. La maggior parte della massa è tuttavia ancora concentrata nei planetesimali di piccole dimensioni. Secondo i calcoli effettuati da Safronov nel 1962 si dovrebbe prevedere che i corpi di grandi dimensioni siano in numero troppo limitato per produrre un aumento della velocità relativa e superare lo smorzamento dovuto alla collisione.

Il modo in cui la situazione si evolve a partire da questo momento è decisivo. Per corpi con un diametro dell'ordine degli otto chilometri, sembra che la velocità relativa aumenti. Finché le basse velocità originarie si mantengono, i primi planetesimali che raggiungono un diametro di 100 chilometri possono avere un vantaggio insuperabile rispetto ai loro vicini più piccoli. In particolare, un incontro fra un planetesimale piccolo e uno di maggiori dimensioni, a una velocità ben al di sotto della velocità di fuga, è in grado di perturbare la traiettoria del planetesimale più piccolo al punto da provocarne la collisione con quello di maggiori dimensioni, anche se apparentemente i due non si trovavano in rotta di collisione. Questo effetto può essere definito «cattura gravitazionale». Allo scopo di prevedere collisioni va tenuto presente che questo effetto aumenta la sezione trasversale effettiva dell'area di influenza del corpo di maggiori dimensioni, oltre la normale dipendenza geometrica dal quadrato del raggio, portandola a valori prossimi alla quarta potenza del raggio.

Se l'effetto di cattura gravitazionale è abbastanza efficace, i primi corpi di grandi dimensioni che si formano catturano i corpi piccoli e quindi impediscono la formazione di corpi di dimensioni intermedie. La distribuzione delle dimensioni dei planetesimali ha quindi due picchi, uno corrispondente ai corpi di grandi dimensioni e l'altro relativo ai corpi piccoli che non sono stati ancora catturati. I corpi grandi si muovono in orbite quasi circolari perché non è avvenuto nulla che abbia aumentato l'eccentricità delle orbite. In verità l'impatto dei piccoli planetesimali su quelli grandi tende a rendere più circolari le orbite di questi ultimi.

Se questo processo continua, i planetesimali grandi acquisiscono la maggior parte della massa del sistema. A questo punto potrebbe essere troppo tardi perché si formi un sistema planetario come quello solare. I corpi grandi si muovono in orbite che non si intersecano, lungo le quali catturano gli ultimi piccoli planetesimali. Il risultato finale è un sistema di circa 5000 pianeti minori, ciascuno con un diametro di circa 1000 chilometri. Si potrebbe pensare che il sistema abbia una seconda possibilità di formare un gruppo di pianeti di tipo terrestre. Forse l'interazione gravitazionale a lungo raggio di migliaia di pianeti minori potrebbe provocare un aumento dell'eccentricità delle orbite al punto da far riprendere l'accrescimento. Tuttavia questa non sarebbe l'evoluzione lineare considerata nel modello di Safronov.

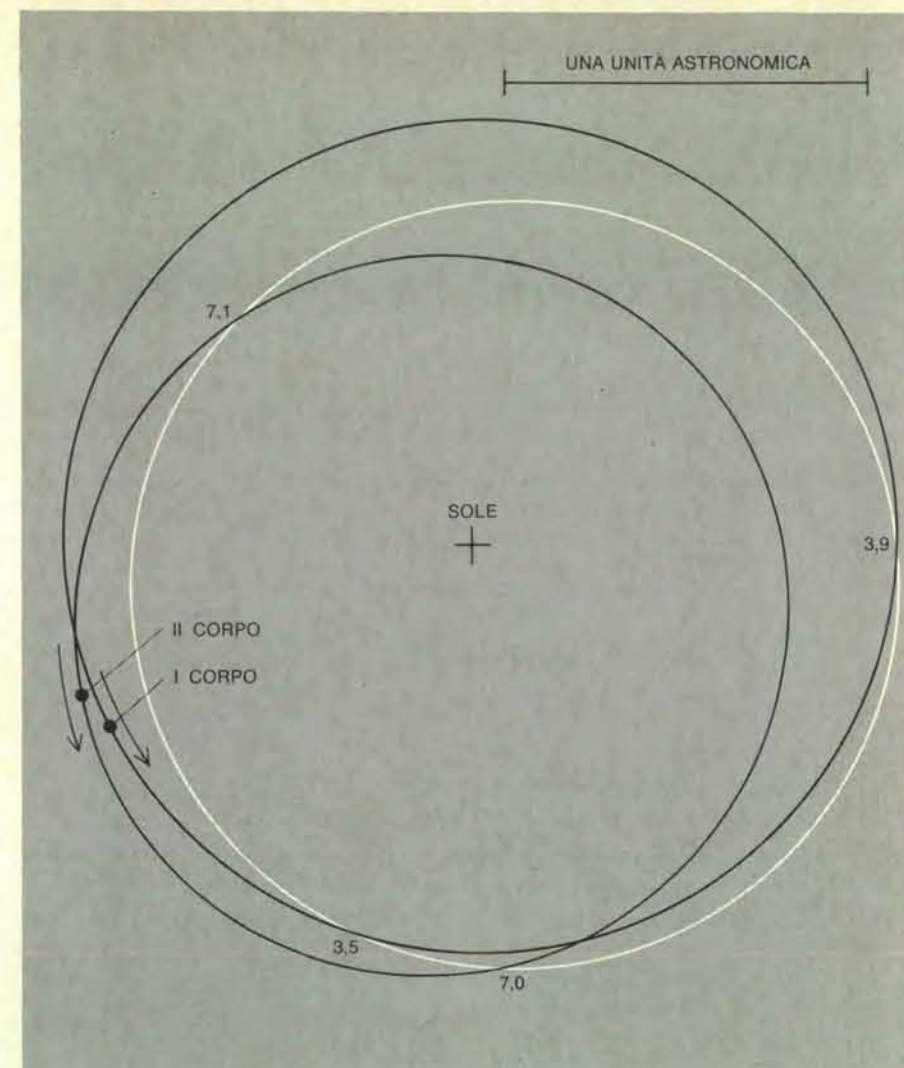
Le interazioni fra due planetesimali sono gli eventi simulati che fanno sì che uno sciame di planetesimali dia origine al sistema solare interno. In questo caso sono state rappresentate le traiettorie di due planetesimali che sono sul punto di interagire. Ciascun corpo ha una massa di 10^{27} grammi e un raggio di circa 4000 chilometri. L'orbita di un terzo planetesimale *(in negativo)* rappresenta il moto degli altri corpi dello sciame. La velocità di ciascun corpo che interagisce con l'orbita del terzo planetesimale è data in chilometri per secondo nei punti d'intersezione delle orbite. Due possibili risultati di interazioni fra due planetesimali sono illustrati nelle figure successive.

Un'altra difficoltà che si ha con il modello di Safronov (e anche con il lavoro di Greenberg) è che esso tratta i planetesimali come se si muovessero lungo percorsi rettilinei come le molecole nella teoria cinetica dei gas. In particolare, Safronov ha esaminato i planetesimali in un'area circostante un planetesimale dato. Da questa posizione di vantaggio le velocità relative dei planetesimali possono essere descritte come se i planetesimali fossero molecole di un gas chiuse in una scatola. Questo metodo però preclude di considerare adeguatamente qualsiasi fenomeno che dipenda dalla distanza dei planetesimali dal Sole.

Il dubbio fondamentale relativo alle semplificazioni implicite nel modello di Safronov è fino a che punto sia valida, in un sistema eliocentrico, l'idea di un equilibrio fra aumento di velocità provocato da interferenze reciproche e diminuzione di velocità dovuta a collisioni. Si potrebbe pensare che l'idea non sia valida. Dopo tutto la velocità di un corpo in un orbita eliocentrica dipende dalla sua distanza dal Sole. Questo fatto introduce una relazione fra distribuzione radiale dei planetesimali e loro velocità relative, che non ha corrispondenze nella teoria cinetica dei gas.

Ho compiuto studi al calcolatore che possono aiutare a risolvere questo dubbio. Lo studio ha inizio con una sciame di 100 corpi di uguale massa. All'inizio i planetesimali seguono orbite eliocentriche di bassa eccentricità. I cambiamenti orbitali derivanti da incontri e collisioni sono di piccola entità, ma numerosi. Ciascun cambiamento viene calcolato e le nuove orbite determinate. Le velocità relative e le eccentricità tendono prima ad aumentare rapidamente, ma successivamente il tasso di aumento diminuisce e infine si stabilizzano su valori proporzionali alla velocità di fuga. Questo è lo stesso effetto trovato anche da Safronov. Se si scelgono orbite iniziali con elevata eccentricità i cambiamenti procedono in direzione opposta: la velocità relativa è elevata all'inizio e successivamente diminuisce, stabilizzandosi sugli stessi valori ottenuti con lo sciame a bassa eccentricità.

Il valore di velocità allo stato stazionario emerso da questi calcoli è circa due terzi del valore previsto dalla teoria di Safronov. Ciononostante l'accordo approssimativo è soddisfacente. La relazione fra velocità relativa e velocità di fuga è confermata, così come la tendenza del sistema ad autoregolarsi in modo che la velocità relativa abbia valori favorevoli all'aggregazione dei corpi.

Questi calcoli tuttavia fanno di più che confermare risultati precedenti. Col procedere dei calcoli, uno sciame inizialmente limitato a un piccolo intervallo di distanze dal Sole si estende fino a includere un intervallo maggiore di distanze. Ciascun planetesimale concorre a determinare questo aumento, in quanto si muove in un intervallo di valori crescenti del raggio dell'orbita, qualche volta vicino a un estremo, qualche volta vicino all'altro, ma più frequentemente mantenendosi su valori intermedi. In effetti il planetesimale si muove casualmente a una certa distanza dal Sole. L'orbita vagante di ciascun planetesimale aumenta le possibilità che corpi distanti entrino in rotte di collisione.

La collisione mancata fra due planetesimali perturba le loro orbite cosicché ciascun corpo segue una nuova traiettoria ellittica attorno al Sole. La velocità del I corpo rispetto al corpo che non è stato coinvolto in questa interazione subisce una diminuzione; quella del II corpo invece subisce un incremento. (Le velocità rispetto al III corpo sono importanti perché difficilmente i due corpi che hanno già interagito verranno nuovamente coinvolti.) In media le collisioni mancate determinano un aumento delle velocità relative dei planetesimali. La collisione mancata di cui viene rappresentato l'esito in questo grafico porta i due planetesimali che interagiscono da una posizione di massimo avvicinamento a una distanza pari a tre volte il raggio di ciascun corpo.

sione. Non bisogna ritenere che l'eccentricità delle orbite sia l'unico fattore che provoca collisioni fra planetesimali. Sebbene l'effetto non elimini la possibilità che si formino pianeti minori troppo piccoli e troppo numerosi, sicuramente tende a diminuirla.

Finora questi calcoli sono stati limitati al caso di corpi con uguale massa. Sono stati inoltre limitati alla ricerca della distribuzione di velocità allo stato stazionario in uno sciame di planetesimali che possono collidere ma non aggregarsi. (Il calcolatore calcola nuove orbite per corpi che collidono, ma non permette loro di aggregarsi.) Attualmente si sta preparando l'estensione dei calcoli al caso di planetesimali con masse diverse.

Uno dei problemi principali consiste nella necessità di capire le circostanze che inducono uno sciame di planetesimali a evolvere esattamente in quattro corpi - i pianeti interni - piuttosto che in un numero diverso. Anche in questo caso si ha più di un processo. Il rallentamento dovuto alla collisione tende sempre a determinare l'accrescimento dei pianeti, isolando i corpi di maggiori dimensioni in orbite quasi circolari. Le perturbazioni gravitazionali tendono invece a produrre l'effetto opposto: provocano infatti un aumento dell'eccentricità dell'orbita e tendono a spostare le orbite stesse radialmente, favorendo le collisioni e l'accrescimento. Prima o poi, quando i corpi di maggiori dimensioni si trovano ormai su orbite che non si intersecano e tutti i corpi più piccoli che attraversano queste orbite sono stati catturati, prevale la tendenza verso orbite quasi circolari. A meno che perturbazioni gravitazionali a lungo raggio non vengano introdotte nella simulazione, non resta alcuna influenza che possa determinare l'intersezione delle orbite dei corpi di

grandi dimensioni. Il problema è quanti di questi corpi grandi esistano.

Larry P. Cox del Massachusetts Institute of Technology ha eseguito alcuni calcoli numerici semplificati sulla crescita simultanea di molti pianeti. Nel suo lavoro si assume innanzitutto che le orbite di 100 corpi di uguali dimensioni siano distribuite casualmente in tutta la regione del sistema solare ora occupata dai quattro pianeti interni. I corpi interferiscono gli uni con gli altri, entrano in collisione e si accrescono. Alla fine restano solo corpi in orbite che non si intersecano. Fino a che punto le dimensioni dei «pianeti» e la loro distribuzione nello spazio assomigliano a quelle dei pianeti interni del sistema solare?

Cox ha scoperto che la risposta dipende dal campo di variabilità delle eccentricità iniziali. Quando il campo di variabilità include corpi con orbite pressoché ellittiche, si nota un'evoluzione sorprendentemente simile a quella dei pianeti interni del sistema solare. In una simulazione il numero finale dei pianeti era sei; uno di essi era un planetesimale molto vicino al Sole, che non era mai entrato in collisione con un altro corpo. Gli altri pianeti avrebbero potuto ad esempio essere chiamati Mercurio, Venere, Terra, Luna e Marte.

Una volta scelte delle eccentricità iniziali più basse (in migliore accordo con il lavoro di Safronov), si formano circa 10 pianeti. La dipendenza dall'eccentricità deriva in parte dal fatto che i calcoli di Cox sono limitati al moto bidimensionale. In questo sistema solare «piatto», le orbite dei planetesimali spesso si intersecano. Come conseguenza le probabilità di uno smorzamento legato alla collisione (inteso in senso opposto alla perturbazione gravitazionale) sono maggiori di quanto si verificherebbe in un sistema tridimensionale. Le mie simulazioni tridimensionali confermano questo concetto. Lo smorzamento dovuto alla collisione sarebbe ancora più importante in uno sciame caratterizzato da componenti con dimensioni iniziali differenti. Un calcolo che iniziasse con corpi equidimensionali dovrebbe perciò introdurre in un modo o nell'altro un numero di collisioni superiore in relazione alle collisioni mancate. Esperimenti di questo tipo sono in corso di studio.

La comprensione definitiva del processo di formazione dei pianeti da planetesimali richiederà lo studio di un altro problema: la formazione dei pianeti nell'ambito di un mezzo gassoso interstellare. Dopotutto la massa del sistema solare è formata quasi interamente da gas altamente volatili quali idrogeno ed elio, i costituenti principali del Sole, di Giove e Saturno. Questi gas pervadono la nebulosa solare primitiva e il loro ruolo è già stato considerato nell'ipotesi che spiega come planetesimali del diametro di un chilometro potrebbero essersi formati dall'aggregazione di granelli di polvere cosmica.

La collisione fra due planetesimali determina la formazione di un singolo corpo di maggiori dimensioni che segue un'orbita quasi circolare. Nella figura è indicata la velocità del corpo di maggiori dimensioni rispetto al corpo non coinvolto nella collisione. In media la collisione ha l'effetto di ridurre le velocità relative dei planetesimali. Inoltre i planetesimali tendono a diventare corpi isolati; infatti, dato che seguono orbite di forma più vicina a quella circolare, è minore la probabilità che vengano a incrociare orbite di altri corpi e quindi a entrare in collisione con essi.

Poiché idrogeno ed elio non sono i componenti principali degli attuali pianeti di tipo terrestre, devono essere sfuggiti da questa regione del sistema solare. Forse il vento solare primordiale era più simile a un uragano solare. In ogni tipo di evento la presenza dei gas deve avere influenzato almeno i primi stadi di aggregazione dei pianeti. Forze addizionali avrebbero successivamente agito sui planetesimali allorché questi si aprirono un varco attraverso un mezzo gassoso resistente; l'entità di queste forze dipende dalla distribuzione del mezzo, che a sua volta dipende dal campo gravitazionale dei planetesimali e dei pianeti, non appena questi si formano. Le simulazioni sono quindi più complesse di quelle che ignorano la presenza di gas. Nei modelli analitici di Chushiro Hayashi e dei suoi colleghi dell'Università di Kyoto è stata inclusa anche la natura del mezzo. Il loro lavoro considera non solo i primi stadi della crescita dei planetesimali (gli stadi descritti da Greenberg), ma anche gli stadi successivi durante i quali si formano corpi con dimensioni pari o superiori a quelle della Luna.

La resistenza aerodinamica provocata dal gas varia in funzione della superficie del corpo che si apre un varco attraverso di essi. Il cambiamento di velocità che deriva da un dato valore di resistenza aerodinamica, è tuttavia minore nel caso di una massa elevata rispetto a una massa di piccola entità, perché il corpo a massa maggiore ha anche una maggiore inerzia. L'effetto di un mezzo gassoso è quindi più importante per corpi di piccole dimensioni. L'effetto su un corpo sufficientemente piccolo è tale che la sua orbita diventa quasi circolare. (Un corpo che segue una traiettoria circolare incontra la minima resistenza da parte dei gas.) Inoltre il corpo si muove a spirale verso il Sole. Per un certo periodo di tempo questo moto a spirale permette ai planetesimali di entrare in collisione e di accrescersi. Pertanto la massa più elevata di questi corpi in crescita diminuisce l'influenza della resistenza. Tuttavia, man mano che i corpi aumentano di massa, diminuisce il loro moto a spirale.

Il risultato sembrerebbe costituito da uno sciame di corpi che percorrono orbite isolate e quasi circolari. Questa prospettiva tuttavia trascura le perturbazioni gravitazionali tra i planetesimali, le quali aumentano in seguito all'accrescersi dei corpi. Le perturbazioni aumentano anche la velocità dei planetesimali rispetto al mezzo gassoso e quindi aumenta anche la resistenza aerodinamica; in realtà l'aumento è approssimativamente proporzionale al quadrato della velocità. Un planetesimale in un mezzo gassoso potrebbe quindi continuare a muoversi a spirale e collidere con altri corpi fino ad avere un diametro superiore a 1000 chilometri.

Nelle simulazioni di Hayashi corpi con un diametro di 2000 chilometri si sono formati entro 10 000 anni. Non è chiaro che cosa sarebbe accaduto in seguito. A dire il vero il mezzo gassoso ha una piccola influenza sui primi corpi di grandi dimensioni che si formano. Essi sono isolati l'uno dall'altro e si trovano su orbite a bassa eccentricità, mentre il resto della massa dello sciame è formata da corpi più piccoli che hanno orbite eccentriche e a spirale. Se questi corpi dovessero collidere e aggregarsi, originando planetesimali con diametro compreso fra 100 e 2000 chilometri, le perturbazioni gravitazionali con i rimanenti planetesimali provocherebbero un aumento della loro velocità rispetto al mezzo gassoso. L'aumento di resistenza aerodinamica renderebbe più pronunciato il loro moto a spirale. È plausibile che in questo modo i planetesimali rimasti possano migrare in senso radiale fino a intersecare le orbite dei corpi di dimensioni maggiori, gli embrioni dei futuri pianeti. Se, tuttavia, i planetesimali in questo moto di migrazione dovessero crescere troppo, potrebbero diventare loro stessi gli embrioni dei pianeti e seguire orbite isolate quasi circolari. Il risultato potrebbe essere un gruppo formato da un numero troppo elevato di pianeti troppo piccoli.

È intanto provato che è molto difficile calcolare le probabilità che questi eventi hanno di verificarsi. Hayashi ha previsto tuttavia la composizione dell'atmosfera che si formerebbe attorno a ciascun nuovo pianeta di tipo terrestre, dovuta alla cattura di gas della nebulosa solare. L'atmosfera contiene anche una certa quantità di gas inerti come argo 36; la concentrazione di questi gas può essere attualmente misurata nelle rocce terrestri. Potrebbe quindi essere possibile conoscere se la Terra si è formata entro un mezzo gassoso basandosi non solo su teorie e simulazioni, ma anche su una più diretta raccolta di dati.

La velocità media di incontro in uno sciame di planetesimali in un dato momento è data dall'equilibrio di due effetti: le collisioni mancate tendono ad aumentare le velocità relative, mentre le collisioni tendono a diminuirle. Il grafico presenta i risultati dei calcoli delle velocità eseguiti per 100 corpi. A ciascun corpo sono stati assegnati una massa di 9×10^{24} grammi e un raggio di 830 chilometri. Le orbite iniziali sono state scelte con bassa eccentricità. Dopo ogni collisione avvenuta o mancata il calcolatore ha ricalcolato le orbite dei corpi partecipanti all'interazione. Il grafico indica la velocità media di incontro sotto l'influenza di collisioni avvenute e mancate. Quando gli effetti di ambedue i tipi di interazioni vengono combinati, la velocità si stabilizza su un valore d'equilibrio pari a 1,6 chilometri per secondo. I calcoli sono stati eseguiti dall'autore.

La maggior parte dei geologi vorrebbero probabilmente scoprire le caratteristiche della Terra ai primordi direttamente piuttosto che basarsi su problemi di fisica e modelli fatti al calcolatore; tuttavia dovranno portare pazienza. Perfino nello stadio attuale del lavoro teorico è possibile, però, dimostrare che fra i tanti modelli ideati, relativi allo stato iniziale della Terra, alcuni sono più verosimili di altri. Il migliore esempio riguarda la temperatura iniziale della Terra. Discussioni riguardo al fatto che la Terra si sia formata a temperatura molto elevata e fosse quindi quasi completamente fusa, o se sia sempre stata un pianeta abbastanza freddo e solido, si sono protratte per generazioni. Il fatto più caratteristico è che ogni teoria attuale sulla formazione dei pianeti prevede una temperatura iniziale elevata.

Innanzitutto vanno considerate le simulazioni relative alla formazione dei pianeti interni da planetesimali, in assenza di gas. Durante i primi stadi di accrescimento in assenza di gas, alcuni dei planetesimali con dimensioni dell'ordine del chilometro formano corpi con un diametro compreso fra 100 e 1000 chilometri a spese dei corpi circostanti più piccoli. Questi corpi si formano in quella regione dello spazio che corrisponderà alla parte del sistema solare occupata dai pianeti interni. Questi corpi possono rappresentare meglio di tutti gli altri gli embrioni dei futuri pianeti. Secondo la descrizione di Safronov «tutti i planetesimali si sono formati nello stesso modo». Si può dimostrare con dei calcoli che solo pochi planetesimali riescono a raggiungere una velocità tale da sottrarsi al sistema solare. Inoltre il sistema solare odierno è vuoto a parte la presenza dei pianeti e dei loro satelliti. Quindi quasi tutti i planetesimali diventano parte dei pianeti, il cui stadio finale di aggregazione è caratterizzato da impatti a elevata velocità fra corpi molto grandi.

Come osservò nel 1918 Joseph Barrell, geologo della Yale University, l'impatto di un corpo di questo tipo cederebbe gran parte dell'energia della collisione al pianeta in corso di accrescimento. In particolare i frammenti della collisione diverrebbero tutti parte del pianeta in accrescimento; la loro energia cinetica verrebbe quindi catturata. Inoltre il calore generato dalla collisione avrebbe bisogno di milioni di anni per attraversare 1000 chilometri di roccia. (Anche qualche metro di roccia può conservare un blocco di ghiaccio durante tutta l'estate.) In effetti il calore verrebbe seppellito con i detriti della collisione. Stranamente l'unica energia che potrebbe sfuggire alla collisione immediatamente potrebbe essere la relativamente piccola quantità di calore dispersa nello spazio, prima della loro caduta, dai frammenti gettati lontano dalla collisione. Calcoli relativi agli effetti termici dell'impatto indicano che corpi in via di accrescimento, di dimensioni non supe-

Il fenomeno di cattura gravitazionale fa sì che un corpo di maggiori dimensioni catturi uno più piccolo anche se questo non sembrava su una rotta di collisione. Nel disegno in alto la velocità di due piccoli corpi («proiettili») rispetto a un corpo di dimensioni maggiori («bersaglio») è pari a 1,45 chilometri per secondo. Uno dei corpi piccoli collide con il bersaglio. In assenza del campo gravitazionale del bersaglio, mancherebbe il centro del bersaglio di una distanza pari a 4,1 volte il raggio del bersaglio stesso. Il secondo corpo di piccole dimensioni avrebbe mancato il bersaglio di una distanza pari a 10 volte il raggio. Invece la sua traiettoria rispetto al bersaglio è stata trasformata in una iperbole e arriva a una distanza pari a cinque volte il raggio. Nel disegno in basso la velocità di ciascun proiettile è 5,8 chilometri per secondo. La forza di attrazione gravitazionale risulta inferiore. Nelle simulazioni relative alla formazione dei pianeti interni del sistema solare il fenomeno di cattura gravitazionale può dar luogo a una disposizione finale con un numero elevato di pianeti relativamente piccoli che hanno catturato tutti i planetesimali a disposizione e che si muovono secondo orbite indipendenti. Il bersaglio nella figura ha una massa di 10^{27} grammi e un raggio di 3970 chilometri. Si ipotizza che ciascun proiettile risulti abbastanza piccolo da non perturbare gravitazionalmente il bersaglio.

riori a quelle della Luna, subirebbero una parziale fusione, mentre corpi con dimensioni analoghe a quelle della Terra sarebbero parzialmente fusi dappertutto.

L'accumulo di materiale roccioso all'interno di un mezzo gassoso, implica inoltre collisioni fra planetesimali di grandi dimensioni. Hayashi ha dimostrato inoltre che, in presenza di un mezzo gassoso, ciascun pianeta in via di accrescimento attrae inizialmente attorno a sé un'atmosfera massiccia. La compressione dell'atmosfera dovuta alla gravità del materiale roccioso che si accumula determinerebbe di per sé un aumento della temperatura superficiale superiore al punto di fusione della roccia, ben prima della formazione definitiva della Terra.

L'unica ipotesi alternativa alla formazione della Terra per accumulo di planetesimali è lo sviluppo di una notevole instabilità gravitazionale entro la nebulosa solare, tale per cui la Terra si aggregherebbe a causa del suo stesso peso. A. G. W. Cameron del Center for Astrophysics dell'Harvard College Observatory e dello Smithsonian Astrophysical Observatory ha dimostrato che anche questo processo implica una elevata temperatura iniziale.

Per parecchi decenni molti studiosi tra cui Harold C. Urey, William W. Rubey e A. P. Vinogradov hanno fornito prove a supporto della teoria opposta. Uno dei più recenti sostenitori di questa opinione è stato A. E. Ringwood dell'Australian National University. Il concetto base della teoria opposta considera la Terra come una grande raffineria chimica. Se una porzione del suo interno raggiunge una data temperatura, alcune sostanze si concentrano nella roccia fusa mentre altre restano separate costituendo un residuo solido. Segregazioni chimiche di questo tipo originano i cristalli di singoli minerali ogniqualvolta la roccia fusa risolidifica. Le segregazioni chimiche associate ai processi di fusione e risolidificazione sono il motivo principale per cui la composizione della crosta continentale è diversa da quella della crosta oceanica. Posto che la Terra ai primordi sia stata più o meno fusa, si potrebbe dubitare che la segregazione che ha prodotto i continenti si sia verificata in tempi così remoti. In questo caso la crosta continentale dovrebbe essere più estesa di quanto non sia in realtà e inoltre non dovrebbe più formarsi nuova crosta.

Perché dunque la differenziazione chimica della crosta continua ancora oggi? La domanda merita una risposta valida, che per il momento tuttavia non può che essere vaga. Il calore generato attualmente all'interno della Terra dal decadimento degli isotopi radioattivi migra verso la superficie soprattutto a causa di fenomeni di convezione che si verificano nel mantello terrestre. Si ritiene che solo una piccola parte della roccia in cui si verificano i moti convettivi sia allo stato fuso. Si sa anche che un basso grado di fusione favorisce la segregazione chimica. Per esempio gli ioni di maggiori dimensioni sfuggono più facilmente da un reticolo cristallino parzialmente fuso. Se la roccia fonde interamente, la sua massa tende invece a rimanere in questo stato.

È stato calcolato che l'impatto dei planetesimali avrebbe prodotto un aumento di temperatura della Terra in via di accrescimento, a una velocità 1000 volte superiore all'attuale riscaldamento radioattivo. Al momento della formazione della Terra tuttavia i moti convettivi dovevano essere molto più rapidi. Una grossa parte di roccia doveva essere allo stato fuso ed è plausibile che il flusso termico nella roccia fusa fosse altrettanto importante del moto convettivo allo stato solido. Si può immaginare che la velocità di convezione fosse rapida, il grado di segregazione chimica basso e la densità delle rocce in via di solidificazione in corrispondenza della superficie non fosse molto inferiore alla densità del sottostante mantello. Diventa quindi concepibile che la crosta appena formatasi venisse semplicemente riassorbita all'interno. In altre parole, forse la tettonica terrestre è un processo lento e la quantità di materiale fuso deve essere bassa per permettere alle rocce a bassa densità che caratterizzano i continenti di venir separate dal circostante materiale e diventare stabili.

Un'altra importante entità geologica che attualmente può venire spiegata in base alle vigenti teorie è il tempo di formazione della Terra. La formazione della Terra per un processo legato all'instabilità gravitazionale all'interno della nebulosa solare richiede circa 100 000 anni. La formazione della Terra da planetesimali in assenza di un mezzo gassoso richiede 100 milioni di anni. Tuttavia la metà della massa terrestre si può aggregare in 20 milioni di anni circa. Sebbene la formazione della Terra da planetesimali all'interno di un mezzo gassoso sia capita meno chiaramente, la presenza di tale mezzo non cambierebbe di molto la scala del tempo.

In contrasto con queste previsioni, le concentrazioni relative dei vari isotopi di piombo nei giacimenti minerari attualmente presenti sulla Terra suggeriscono (basandosi sulla semplice interpretazione dei dati e assumendo che la temperatura della Terra primordiale fosse elevata) che il processo di formazione abbia richiesto un lungo periodo di tempo. A dire il vero esse suggeriscono che dopo 100 milioni di anni il processo di aggregazione fosse a metà. Tutti i modelli teorici relativi alla formazione della Terra non sono d'accordo tuttavia con la semplice interpretazione dei dati isotopici. Forse un'interpretazione più complessa, ma non per questo meno credibile dei dati, porterebbe infine a una più ampia conoscenza della storia chimica della Terra ancora in atto. Una scienza per essere valida deve essere caratterizzata da questo scambio fra osservazione e teoria.

È probabile che fino a oggi la maggior parte dei geologi abbia considerato queste teorie relative alla formazione della Terra come eccessive speculazioni o innocente passatempo. Tuttavia oggi stanno diventando uno studio serio, in quanto la maggior parte della ricerca geologica, così come la progettazione di missioni spaziali, richiede conoscenze sulle modalità di formazione della Terra. La piena comprensione di questo processo tuttavia richiederà ancora del tempo.

Programmazione lineare e allocazione di risorse

Strutture astratte simili a cristalli pluridimensionali possono aiutarci a risolvere problemi di programmazione e di amministrazione. Un nuovo algoritmo pone un limite superiore alla complessità di questi problemi

di Robert G. Bland

Consideriamo la situazione di una piccola fabbrica di birra i cui prodotti, e cioè la birra e lo «ale» (bevanda molto diffusa negli Stati Uniti simile alla birra ma più alcoolica), vengano costantemente richiesti dal mercato ma la cui fabbricazione sia limitata da alcune materie prime scarseggianti. Supponiamo che queste ultime siano il mais, il luppolo e il malto d'orzo. La ricetta per un barile di ale richiede un dosaggio degli ingredienti differente da quello per un barile di birra. Lo ale, ad esempio, vuole più malto per barile che non la birra. Inoltre il fabbricante ricava dalla vendita dello ale un profitto di 13 dollari a barile, contro 23 dollari a barile di birra. Soggetto a tali condizioni, come può il fabbricante massimizzare i propri profitti?

Si potrebbe pensare che il programma produttivo migliore sia di impiegare tutte le risorse nella fabbricazione del prodotto che garantisce maggiori profitti, e cioè la birra. Questa scelta tuttavia può non essere la più saggia, dal momento che la produzione di questo tipo di birra può consumare alcune delle risorse disponibili con una rapidità maggiore dello ale. Se ci vogliono 5 libbre (1 libbra = 0,453 chilogrammi) di mais per fare un barile (1 barile = 119,22 litri) di ale e 15 libbre per fare un barile di birra, è possibile produrre il triplo di ale che di birra. Inoltre, producendo solo birra, il fabbricante scoprirà di consumare tutto il mais prima di aver esaurito le scorte di luppolo e di malto. Un uso più vantaggioso delle risorse e quindi un profitto maggiore si possono invece ottenere dalla produzione mista di ale e di birra. Ma definire un simile programma produttivo di ottimo non è un problema banale. È esattamente questo il tipo di problemi la cui formulazione esplicita, analisi e soluzione, sono affidate alle tecniche della programmazione lineare. Quest'ultima è una branca della matematica che rientra nel più ampio campo d'indagine della cosiddetta ricerca operativa, in cui i più disparati metodi di modellistica matematica e di analisi quantitativa vengono applicati alla gestione di grandi imprese e organismi. La programmazione lineare fu messa a punto nell'immediato secondo dopoguerra in risposta a problemi logistici che si erano presentati durante e subito dopo la guerra. Una delle prime pubblicazioni sull'impiego della programmazione lineare conteneva la discussione di un modello del peso totale aerotrasportato a Berlino nel 1948.

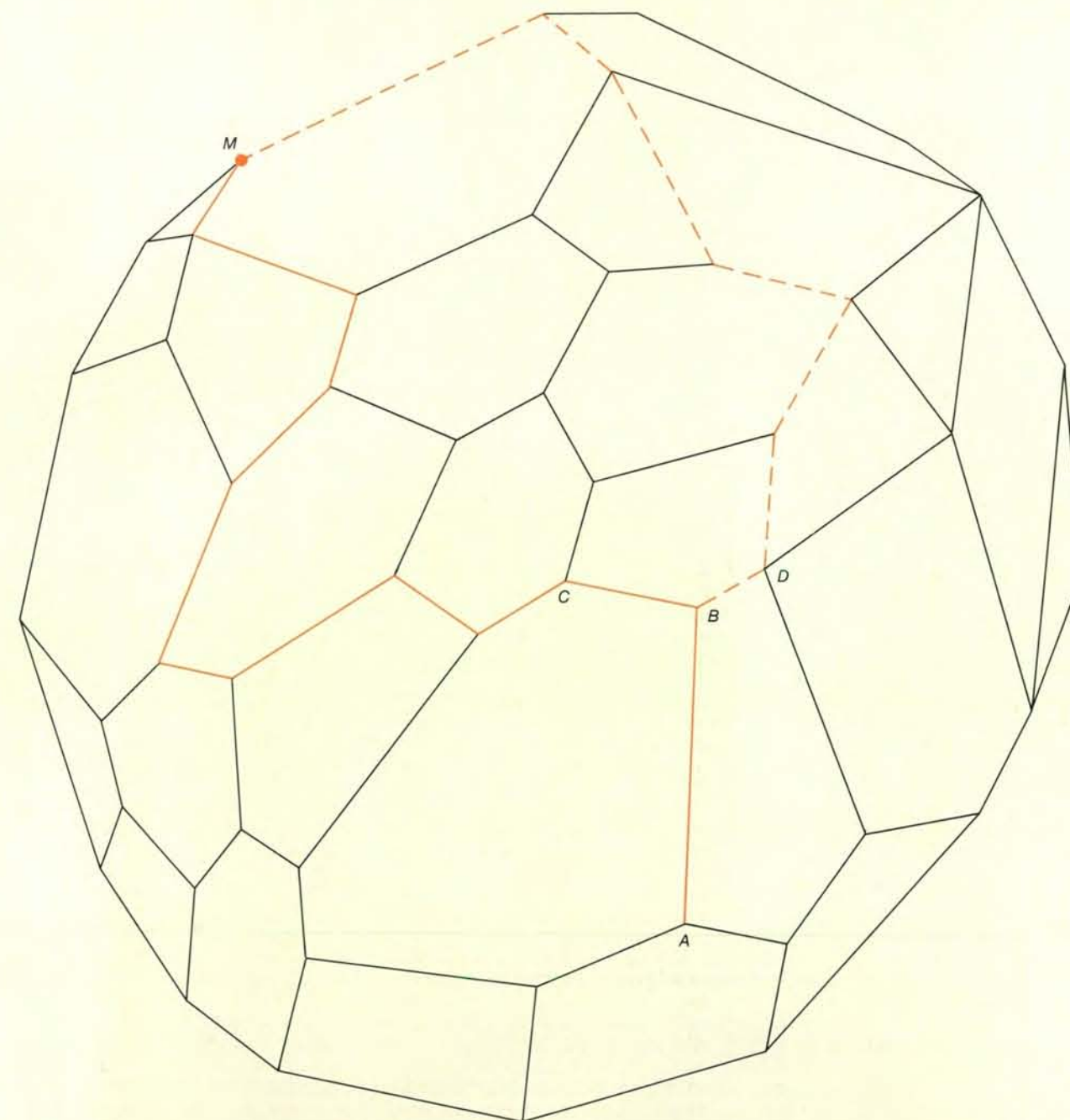
Per quanto l'elaboratore sia uno strumento indispensabile per risolvere problemi di programmazione lineare, il termine «programmazione» è inteso nel senso di stesura di piani e non di formulazione di programmi per elaboratore. L'aggettivo «lineare» si riferisce invece a una proprietà matematica di certi problemi che ne semplifica l'analisi. Nel problema della fabbrica di birra si suppone, per esempio, che il quantitativo di una qualsiasi risorsa necessaria a produrre ale o birra sia proporzionale alla quantità di bevanda prodotta. Raddoppiando cioè la quantità di birra, si raddoppia la quantità di ciascun ingrediente necessario alla sua produzione, e si raddoppia anche il profitto imputabile alla sua vendita. Se si traccia il diagramma cartesiano del consumo di mais in funzione della quantità di birra prodotta con il medesimo, si ottiene una retta. Per applicare le tecniche della programmazione lineare si deve altresì supporre che i prodotti e le risorse siano quantità divisibili, almeno in via approssimativa. Così si potrà produrre mezzo ettolitro di birra con un valore che sarà la metà di quello di un ettolitro intero.

Un tipico ambito dei problemi di programmazione lineare è la ripartizione di risorse scarse su un certo numero di prodotti o attività, alle condizioni di proporzionalità e divisibilità ora descritte. Tali risorse possono essere materie prime, prodotti semilavorati, manodopera, capitali investiti o tempi di lavoro su macchine di grandi dimensioni e molto costose. Un'allocazione di ottimo sarà quella che massimizza una certa misura di benefici o utilità, come il profitto, o ne minimizza una di costi. Nella nostra epoca di produttività decrescente e di risorse in diminuzione sarà certamente utile esaminare una tecnica capace di allocare le risorse con la massima efficienza possibile.

Le proprietà dei problemi di programmazione lineare discendono da alcuni principi elementari di algebra e di geometria. La soluzione efficiente di tali problemi dipende da algoritmi, o procedure completamente specificate passo per passo, in grado di sfruttare nel modo più intelligente questi principi algebrici e geometrici. Questi algoritmi sono altresì di semplice concezione, per quanto complicati possano apparire i dettagli delle loro operazioni. È dall'efficacia e dalla versatilità di un singolo algoritmo, il cosiddetto «metodo del simplesso», che discende in gran parte il rilievo economico della programmazione lineare.

Il metodo del simplesso è stato introdotto nel 1947 da George B. Dantzig, ora alla Stanford University. Il suo valore essenziale consiste nella rapidità, nella ricchezza delle applicazioni e nella possibilità di rispondere a importanti quesiti circa la suscettibilità delle soluzioni a variazioni dei dati in entrata. Per esempio: in che misura deve essere modificato il programma produttivo del fabbricante di birra se si sono modificate le disponibilità di luppolo e la resa economica della birra? Quanto dovrà essere disposto a pagare per ulteriori scorte di risorse che scarseggiano? Quanto dovrà farsi pagare da un altro imprenditore che voglia comprare da lui delle risorse? Il criterio del simplesso può aiutare a decidere se sia meglio acquistare o vendere una macchina, farsi prestare denaro o concedere prestiti e pagare degli straordinari o farne a meno. Con il suo aiuto è inoltre possibile imporre vincoli aggiuntivi e risolvere nuovamente il problema onde esaminarne gli effetti. Esso può, per esempio, dire tempestivamente all'imprenditore quanto costa fornire un servizio che non dà profitto ma che può assicurargli la buona disposizione del cliente.

Il metodo del simplesso si è dimostrato estremamente efficiente nella soluzione di problemi complessi di programmazione lineare con migliaia di vincoli, anche se la sua rapidità ha sempre rappresentato per i teorici un enigma, mancando una spiegazione definitiva del motivo di queste ottime prestazioni. Di fatto, vi sono problemi costruiti dai matematici per i quali il metodo del simplesso è di una lentezza inaccettabile. Ma per ragioni che non sono del tutto chiare, tali problemi non sembrano presentarsi nella pratica.

Alcuni matematici sovietici hanno recentemente messo a punto un nuovo algoritmo per la programmazione lineare che evita, in un certo senso, alcune difficoltà teoriche attribuite al metodo del simplesso. A dimostrazione del significato economico della programmazione lineare oggi, la notizia di questo risultato è apparsa sulle prime pagine dei giornali in tutto il mondo. Sfortunatamente, il nuovo algo-

Il metodo del simplesso è una procedura molto usata per risolvere problemi di programmazione lineare. Esso consente di trovare un'allocazione ottimale delle risorse spostandosi da un vertice all'altro lungo gli spigoli di un politopo, solido tridimensionale le cui facce sono poligoni. Ciascun punto della porzione di spazio in cui è costruito il politopo corrisponde a un particolare programma di allocazione di manodopera, capitale o altre risorse e a ciascun programma corrisponde un beneficio (o costo) netto. Scopo della programmazione lineare è di individuare il programma che presenti il massimo di benefici e il minimo di costi. Il politopo definisce una zona «ammissibile»: tutti i programmi di allocazione delle risorse rappresentati da punti contenuti nel politopo o appartenenti alla sua superficie sono ammissibili, mentre quelli rappresentati da punti esterni al politopo non lo sono a causa della scarsità di risorse. Quando il rapporto fra risorse e costi o benefici è lineare, i valori di massimo e di minimo devono giacere in corrispondenza di uno dei vertici del politopo. L'algoritmo del simplesso passa in rassegna in modo selettivo i vertici, tracciando un percorso *(ABC...M)* lungo gli spigoli del politopo. A ciascun passo del percorso la misura dei benefici o dei costi continua a migliorare fino a raggiungere nel punto *M* il massimo o il minimo. Vi sono spesso numerosi percorsi che vanno dal vertice di partenza *A* a *M*. Il politopo non deve necessariamente essere tridimensionale: anzi comunemente ha migliaia di dimensioni.

L'applicazione di attribuzione cerca di minimizzare il costo dell'adattamento di edifici disponibili per il ripristino a funzioni che debbono essere svolte ciascuna in un solo edificio. Con tre edifici e tre funzioni occorre considerare 3^2 ovvero nove costi per altrettanti interventi di rinnovamento. I costi (valutati in miliardi di lire) sono riportati in una matrice numerica; un'attribuzione ammissibile sceglie un numero per ciascuna riga e ciascuna colonna della matrice. Vi sono 3 × 2 × 1 ovvero sei alternative possibili. Più in generale, per attribuzioni di dimensione *n*-per-*n* il numero delle attribuzioni è *n* fattoriale (scritto *n!*), pari a *n* moltiplicato per tutti gli interi positivi minori di *n*. A causa della rapidità con cui cresce *n!* è praticamente impossibile determinare il costo minimo mediante enumerazione di tutte le possibili attribuzioni.

ritmo, che va sotto il nome di «metodo dell'ellissoide», ha finora poche speranze di superare nella pratica il metodo del simplesso. Il fatto è che, al presente, vi è una curiosa divergenza fra misure pratiche e teoriche di esecuzione dei calcoli.

Anche quando si utilizza un algoritmo efficiente, i costi di messa a punto per la soluzione di un problema molto ampio di programmazione lineare possono essere considerevoli. L'esplicitazione di un insieme reale di circostanze in termini di programmazione lineare - come pure la raccolta e l'organizzazione dei dati che le descrivono - non è impresa da poco. Inoltre la soluzione del problema è possibile solo utilizzando un elaboratore veloce.

Tuttavia i benefici economici della programmazione lineare sono spesso decisivi. Verso la metà degli anni cinquanta, allorché furono messi a punto dei metodi di programmazione lineare per coordinare la miscelatura delle benzine, la Exxon Corporation cominciò a realizzare risparmi dal 2 al 3 per cento sul costo delle operazioni di miscelatura. L'applicazione si estese presto nell'industria petrolchimica al controllo delle ulteriori operazioni di raffinazione, compresi il cracking catalitico, la distillazione e la polimerizzazione. Press'a poco nello stesso periodo altre industrie, e in particolare quella cartaria, quella della distribuzione degli alimenti, l'agricoltura, le acciaierie e l'industria metallurgica, incominciarono ad adottare la programmazione lineare. Secondo una stima di Charles Boudrye della Linear Programming Inc., di Silver Spring, Maryland, i profitti di una cartiera sono aumentati di 15 milioni di dollari grazie all'applicazione della programmazione lineare per determinare l'assortimento dei prodotti.

Oggi vengono offerti da una decina di società «pacchetti» di programmi per elaboratore basati sull'algoritmo del simplesso e circa 1000 clienti ne fanno uso su licenza dei loro estensori. Ciascun cliente paga consistenti diritti mensili, sicché è probabile che egli faccia uso del suddetto metodo in modo regolare. Un numero molto maggiore di organismi ha invece accesso ai «pacchetti» attraverso dei consulenti. Inoltre, sono stati messi a punto dei programmi specializzati per risolvere problemi di flussi nelle reti, i quali possono fornire un servizio ancor più ampio degli algoritmi generici di programmazione lineare.

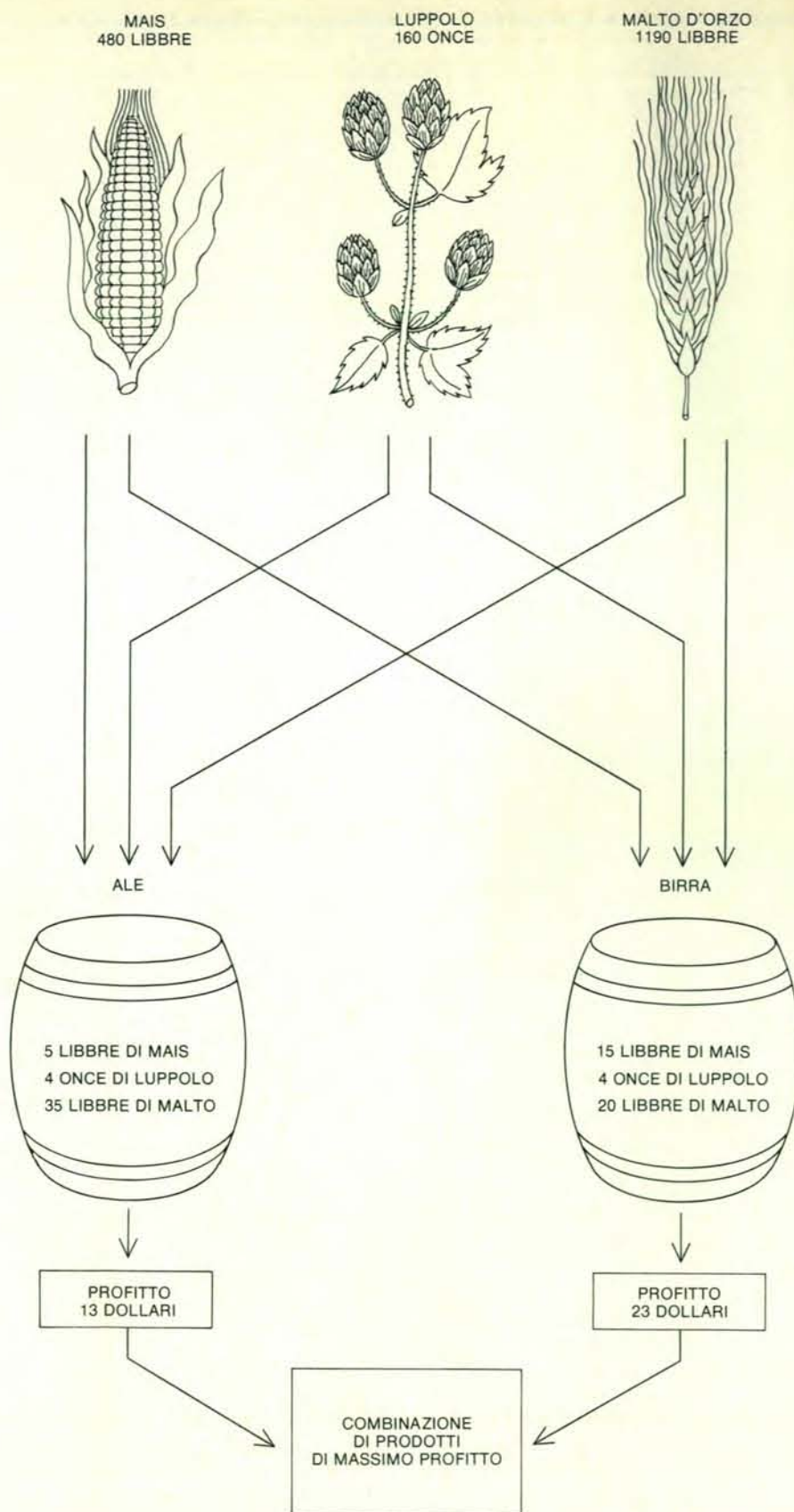
La vasta applicabilità della programmazione lineare è riconoscibile anche in un singolo organismo. La Exxon, per esempio, applica correntemente la programmazione lineare alla determinazione dei tempi nelle operazioni di trivellazione, all'allocazione del grezzo alle diverse raffinerie, alla distribuzione dei prodotti e alla pianificazione delle strategie di mercato. David Smith, del Communication and Computer Sciences Department della Exxon, riferisce che alla programmazione lineare e alle sue estensioni è riservato il 5-10 per cento del carico totale di elaborazione della società. Negli ultimi 20 anni tale quota ha tenuto il passo con le applicazioni generali di elaborazione dell'informazione in rapido sviluppo.

Sebbene il metodo del simplesso sia uno strumento molto potente, esso si basa su idee elementari. Per meglio intenderne alcune è utile esaminare un problema di allocazione dalla struttura particolare, la cosiddetta «applicazione di attribuzione». Consideriamo il caso di una commissione per la progettazione di un'università, con tre edifici disponibili per il rinnovamento strutturale e tre funzioni diverse a cui essi debbono venire adibiti. Immaginiamo che le funzioni siano quelle di laboratorio, biblioteca e campi da tennis coperti, per cui ogni edificio non può assolvere più di una funzione. Le tavole della pagina a fronte indicano il costo di rinnovamento delle nove possibili combinazioni funzione-edificio. In che modo la commissione può minimizzare i costi di rinnovamento strutturale?

Il problema si può risolvere scegliendo

	SI SOTTRAE 3	SI SOTTRAE 4	SI SOTTRAE 8	SI SOTTRAE 2	SI SOTTRAE 9	SI SOTTRAE 7	SI SOTTRAE 0	SI SOTTRAE 5	SI SOTTRAE 1	SI SOTTRAE 6
SI SOTTRAE 2	6	6	12	4	14	11	3	9	4	10
SI SOTTRAE 5	8	12	17	8	16	16	6	13	7	13
SI SOTTRAE 6	11	13	18	9	15	16	8	14	8	14
SI SOTTRAE 4	8	10	15	10	16	13	7	9	7	11
SI SOTTRAE 8	13	15	16	11	19	19	10	16	12	18
SI SOTTRAE 5	9	12	17	8	16	15	7	13	7	11
SI SOTTRAE 3	8	8	13	8	13	12	6	8	4	11
SI SOTTRAE 1	6	8	11	3	14	10	2	8	3	9
SI SOTTRAE 0	5	5	12	2	12	7	2	9	6	12
SI SOTTRAE 7	15	14	17	13	17	18	7	18	10	16

1	0	2	0	3	2	1	2	1	2
0	3	4	1	2	4	1	3	1	2
2	3	4	1	0	3	2	3	1	2
1	2	3	4	3	2	3	0	2	1
2	3	0	1	2	4	2	3	3	4
1	3	4	1	2	3	2	3	1	0
2	1	2	3	1	2	3	0	0	2
2	3	2	0	4	2	1	2	1	2
2	1	4	0	3	0	2	4	5	6
5	3	2	1	1	4	0	6	2	3

Convincere uno scettico che un'attribuzione è ottimale *(illustrazione in alto, caselle in colore)* non comporta la stima dei costi di tutte le attribuzioni ammissibili. In questo caso la matrice dei costi è 10 × 10, con conseguenti 10! ovvero 3 628 800 attribuzioni possibili. Sottraendo un determinato numero da ciascuna entrata di una riga o di una colonna, lo si sottrae anche dal costo totale di ogni possibile attribuzione, con il risultato di lasciare invariati i costi relativi. La ragione è che ciascuna attribuzione sceglie esattamente un numero dalla riga (o colonna) trasformata della matrice. L'insieme di numeri da sottrarre può essere scelto in modo che la matrice di partenza si trasformi in una matrice senza entrate negative e con almeno un'entrata a costo nullo in ciascuna riga o colonna *(in basso)*. Dal momento che nessuna attribuzione può avere un costo totale minore di zero, l'attribuzione ottimale per la matrice trasformata deve presentare un costo totale nullo. L'immediata conseguenza è che anche le corrispondenti entrate nella matrice di partenza sono ottimali. Per generare l'insieme di numeri da sottrarre sono stati escogitati algoritmi efficienti.

Il dilemma del fabbricante di birra rappresenta un esempio di applicazione della programmazione lineare a problemi di ottimizzazione dello smistamento di risorse su diversi canali produttivi. La produzione di birra normale e di «ale» (una sorta di birra a più alta gradazione alcoolica assai diffusa negli Stati Uniti) è limitata dalla scarsità dei tre ingredienti essenziali: mais, luppolo e malto. I livelli di produzione ammissibili sono determinati non solo dal quantitativo totale di ciascun ingrediente disponibile, ma anche dalle proporzioni richieste dai due prodotti. La funzione obiettivo, e cioè la quantità da ottimizzare, è il profitto del fabbricante di birra. Nella programmazione lineare si suppone che tutte le risorse, i prodotti e i benefici siano quantità divisibili: il fabbricante può usare mezza libbra di mais, vendere un quarto di barile di ale e realizzare profitti proporzionali.

tre caselle nella tavola 3 × 3, e precisamente una per riga e una per colonna in modo che a ciascun edificio venga assegnata una funzione e che tutte le funzioni siano coperte. La soluzione ottimale è quella che minimizza la somma dei costi nelle caselle prescelte. La sua individuazione non è difficile, dal momento che le alternative possibili sono poche. Una volta infatti che sia stata segnata una delle caselle della prima riga, nella seconda ne restano libere solo due. Nella terza, poi, la scelta è forzata, dal momento che rimane una sola casella nella colonna non ancora segnata. Vi sono perciò 3 × 2 × 1, ovvero sei modi possibili di operare l'attribuzione. È facile enumerare tutte e sei le possibilità, valutare il costo totale di ciascuna e prescegliere l'attribuzione economicamente più vantaggiosa.

Questo approccio enumerativo, se risolve il problema di una matrice 3 × 3, diventa invece impraticabile per problemi di dimensioni maggiori. Nel caso di 4 edifici e 4 diverse funzioni, il numero delle possibili attribuzioni è 4 × 3 × 2 × 1, cioè 24. Nell'enunciazione generalizzata del problema vi sono n edifici e n funzioni, mentre il numero delle attribuzioni è n fattoriale (scritto $n!$), e cioè n moltiplicato per tutti gli interi da 1 a n-1. Così, per $n = 10$ vi saranno 10!, ovvero più di 3,6 milioni di attribuzioni differenti.

La rapida crescita di $n!$ dissipa qualsiasi entusiasmo per il metodo enumerativo. Immaginiamo di dover risolvere mediante enumerazione un'applicazione di attribuzione 35 × 35 e di avere un elaboratore capace di scegliere fra le attribuzioni possibili, valutare il costo di ciascuna e confrontarla con l'attribuzione con il costo più basso fino a quel momento, il tutto al ritmo di un miliardo di attribuzioni al secondo (un elaboratore in grado di funzionare a tale velocità sarebbe molto più rapido di quelli attualmente disponibili). Anche se il compito di enumerare le 35! attribuzioni fosse suddiviso su un miliardo di elaboratori di questo tipo, dopo un miliardo di anni potrebbe essere completata solo una frazione insignificante dei calcoli necessari.

L'applicazione di attribuzione 35 × 35 non è particolarmente grande. Se si tratta di assegnare del personale a diverse mansioni per minimizzare il costo totale dell'addestramento professionale, n può facilmente essere uguale a 35. Ma vi sono molte applicazioni di attribuzione in cui n è uguale a 1000 e più. Problemi di questo tipo richiedono un procedimento più «astuto» della semplice enumerazione.

Il peso dell'enumerazione verrebbe grandemente ridotto se si potesse evitare di passare in rassegna le attribuzioni che comportano costi maggiori di quelle già controllate. Ciò sarebbe possibile se vi fossero una regola di arresto o un criterio di ottimalità capaci di individuare rapidamente e di primo acchito le attribuzioni di ottimo. Un algoritmo che incorporasse un criterio del genere offrirebbe significativi benefici collaterali. I benefici si possono riassumere in quello che Jack Edmonds, della Università di Waterloo, nell'Ontario, chiama «il principio del supervisore assoluto» e che si potrebbe anche chiamare dello «scetticismo del capo».

Immaginate, dopo noiose enumerazioni, di aver risolto l'applicazione di attribuzione 10 × 10, indicata nell'illustrazione in alto della pagina 95, passando in rassegna tutte le 3 628 800 possibili attribuzioni. L'attribuzione di ottimo è, secondo voi, quella segnata in colore nella figura. Presentate soddisfatti la soluzione al vostro capo, il quale vi guarda bene negli occhi, tira una boccata dal sigaro e vi domanda a bruciapelo: «Ma è proprio sicuro che non ci sia una soluzione più conveniente?». Avete un bel deglutire, visto che l'unico modo di mostrare i meriti della vostra soluzione sembra proprio essere quello di ripetere al capo lo scrutinio di tutte le 3 628 800 possibilità.

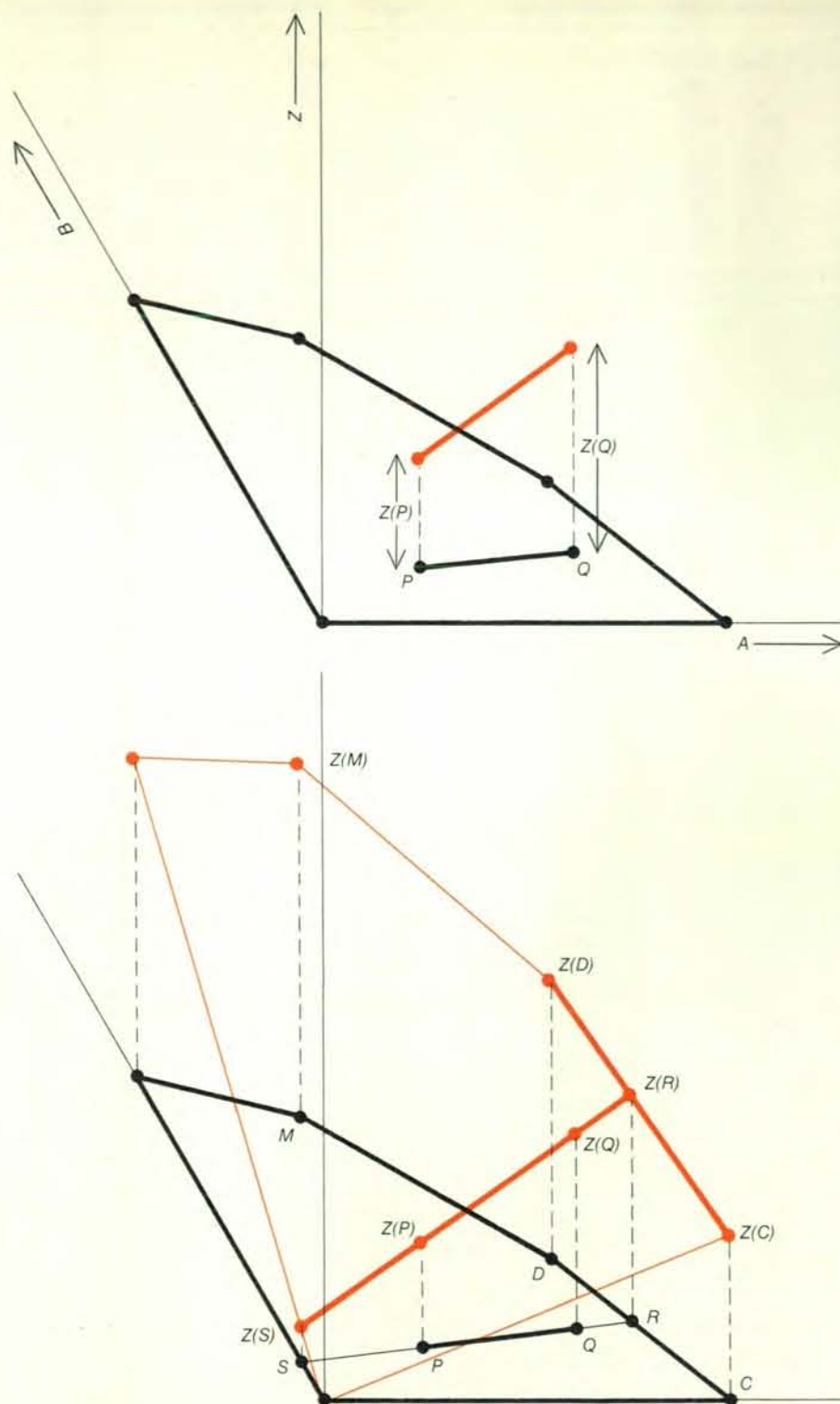
Una regola di arresto è capace di offrire una concisa dimostrazione di ottimalità. Immaginate di presentarvi dal vostro capo non solo con un'attribuzione di ottimo, ma anche con un insieme di numeri da sottrarre alle entrate di ciascuna riga e colonna. Spiegare il modo in cui questi numeri sono ottenuti richiederebbe una discussione particolareggiata dell'applicazione di attribuzione; basterà specificare che l'insieme dei numeri può essere individuato mediante un algoritmo efficiente su un elaboratore. La loro utilità, una volta individuati, appare subito evidente. Si osservi che sottrarre lo stesso numero da ciascuna entrata di una riga o colonna date equivale a sottrarlo dal costo totale di ogni possibile attribuzione; il motivo è che in ciascuna delle attribuzioni possibili si deve scegliere una, e una sola, entrata di ogni riga e colonna. Se, ad esempio, si sottrae 5 a tutte le entrate della sesta riga, tutte le attribuzioni possibili conterranno esattamente un'entrata di 5 punti minore della relativa attribuzione fatta con l'insieme iniziale dei costi. I costi relativi di tutte le attribuzioni possibili rimarranno pertanto immutati. Tale sottrazione può essere ripetuta purché sia sempre applicata uniformemente a ciascuna entrata di una data riga o colonna.

Tramite sottrazione iterata è possibile trasformare la matrice iniziale dei costi nella matrice illustrata nella figura in basso a pagina 97. Questo insieme di costi possiede un'interessante proprietà. Potete cioè segnalare al vostro capo che i costi corrispondenti alle caselle scelte dalla vostra attribuzione sono tutti nulli e che nessuna entrata della matrice è negativa. Poiché la somma dei costi segnalati dalla vostra attribuzione è nulla e non vi sono costi negativi, non esistono possibili attribuzioni a costi inferiori. In breve, avete dimostrato al vostro capo, con poche migliaia di calcoli anziché con decine di milioni, che nessuna attribuzione può essere meno costosa di quella da voi prescelta.

La zona ammissibile nel problema del fabbricante di birra è formata dall'intersezione di cinque semipiani. A ogni punto (A, B) del piano corrisponde un programma produttivo che richiede la produzione di A barili di ale e B barili di birra. I primi tre semipiani rappresentano graficamente tutti i programmi produttivi realizzabili con gli ingredienti a disposizione. Ad esempio, il quantitativo di mais richiesto è $5A + 15B$ (peso in libbre del mais necessario a produrre 1 barile di ale moltiplicato per A più il peso del mais necessario a produrre 1 barile di birra moltiplicato per B). Tale quantità non deve superare le 480 libbre di mais effettivamente disponibili per il fabbricante. Quindi, ogni punto del semipiano dominato dalla retta $5A + 15B = 480$ rappresenta un piano produttivo ammissibile che non richiede più di 480 libbre di mais. In modo analogo si possono costruire i semipiani relativi al luppolo e al malto. Gli altri due semipiani esprimono il fatto che sono ammissibili solo quei programmi che hanno una produzione non negativa.

Per quanto non abbia dimostrato il modo di risolvere un'applicazione di attribuzione o di trovare le costanti di riga o di colonna da sottrarre, il problema dell'attribuzione illustra la necessità di evitare l'enumerazione e la possibilità di farlo mediante una regola di arresto capace di riconoscere una soluzione di ottimo. Si tratta di caratteristiche strutturali degli algoritmi che si possono applicare anche ai problemi di programmazione lineare in generale.

Consideriamo di nuovo il caso del fabbricante di birra i cui due prodotti, lo ale e la birra normale, sono fatti con diverse proporzioni di mais, luppolo e malto. Supponiamo che vi sia disponibilità immediata di 480 libbre di mais, 160 once di lievito e 1190 libbre di malto e che la produzione sia limitata dalla scarsità di quei materiali. Altre risorse consumate nel processo produttivo, come l'acqua, il lievito, la manodopera e l'energia, si suppone esistano in abbondanza: sebbene possano con i loro costi condizionare le intenzioni del fabbricante a produrre ale o birra, non limitano comunque direttamente la capacità produttiva. Supponiamo che la fabbricazione di ciascun barile di ale richieda 5 libbre di mais, 4 once di luppolo e 35 libbre di malto. Supponiamo ancora che tutto lo ale e tutta la birra che si riesce a produrre possano essere venduti ai prezzi correnti, con un profitto di 13 dollari per barile di ale e 23 dollari per barile di birra.

La scarsità di mais, luppolo e malto limita i livelli produttivi ammissibili. Così, per esempio, sebbene vi siano abbastanza luppolo e malto per fabbricare più di 32 barili di birra, la produzione di quest'ultima esaurirebbe la disponibilità di mais non consentendo un'ulteriore produzione di birra e impedendo del tutto quella di ale. Un altro programma produttivo ammissibile non prevede birra, ma solo 34 barili di ale, esaurendo in tal modo tutte le 1190 libbre di malto. La prima alternativa sembra da preferirsi alla seconda. Il profitto realizzato dal primo programma è di 32 × 23, ovvero 736 dollari contro i soli 34 × 13, ovvero 442 dollari, del secondo programma.

Il valore della funzione obiettivo per ogni punto *P* della zona ammissibile può essere rappresentato graficamente dalla distanza *Z (P)* al di sopra o al di sotto di *P*, misurata sull'asse *z* (le distanze lungo l'asse *z* sono state tracciate a una scala differente da quella per le distanze sul piano della zona ammissibile). Si può pensare alla distanza *Z (P)* come a un punto dello spazio tridimensionale. Se la funzione è lineare, il grafico dei valori assunti dalla funzione lungo una retta compresa nella zona ammissibile è anch'esso una retta *(grafico in alto)*. Per ogni punto interno a tale zona, esiste una retta passante per *P* che interseca il contorno della zona in due punti, per esempio *R* e *S* *(grafico in basso)*. Se il segmento di retta nello spazio che unisce *Z (R)* e *Z (S)* non è parallelo al piano della zona ammissibile, la funzione obiettivo deve assumere il valore massimo lungo il segmento a uno dei suoi estremi, per esempio *Z (S)*, che corrisponde al punto *S* sullo spigolo della zona ammissibile. Il grafico della funzione obiettivo lungo uno spigolo è una retta e assume anch'esso il valore massimo a uno degli estremi, diciamo *Z (D)*, che corrisponde a un vertice della zona ammissibile. Vi è sempre un punto sullo spigolo che domina un punto interno dato, e c'è sempre un vertice che domina un punto lungo uno spigolo. Per trovare il valore massimo *Z (M)* della funzione obiettivo è necessario esaminare solo i vertici. Se la zona ammissibile è bidimensionale, la funzione obiettivo forma un piano la cui altezza massima è raggiunta in un vertice.

Esistono altri programmi produttivi migliori di questi due. Così, 6 barili di ale e 30 barili di birra sfruttano tutte le 480 libbre di mais, 154 delle 160 libbre di luppolo e 810 delle 1190 libbre di malto, garantendo un profitto di $(6 \times 13) + (30 \times 23)$, ovvero 768 dollari. Con altri programmi si ottengono profitti anche maggiori. In tal caso, l'enumerazione di tutte le possibilità non è soltanto impraticabile, come nel caso dell'applicazione di attribuzione; è addirittura impossibile. Vi è un numero infinito di programmi produttivi che risolvono il problema del fabbricante di birra e ciascuno di essi va sotto il nome di soluzione ammissibile. Vi è fortunatamente solo un insieme limitato di soluzioni ammissibili su cui val la pena di soffermare l'attenzione, l'insieme delle cosiddette «soluzioni estremali».

L'importanza delle soluzioni estremali risulta evidente se si rappresenta graficamente l'insieme di tutte le soluzioni ammissibili come un insieme di punti del piano, insieme che costituisce appunto la cosiddetta zona ammissibile. Indichiamo con A il numero di barili di ale fabbricati secondo tutte le possibili strategie produttive e con B il numero di barili di birra. A e B sono note nella programmazione lineare come variabili decisionali e possono essere associate alle coordinate del piano. Tutti i punti del piano si possono indicare con una coppia di coordinate (A, B), corrispondenti altresì a un particolare insieme di livelli produttivi.

Essendo impossibili dei livelli produttivi con valori negativi, si può immediatamente confinare la zona ammissibile al quadrante superiore destro del piano, dove A e B sono entrambi non negativi. In che modo la scarsità di malto influisce sulla produzione? Dal momento che occorrono 35 libbre di malto per ogni barile di ale e 20 libbre per ogni barile di birra, il quantitativo totale di malto necessario a fabbricare A barili di ale e B barili di birra è $36A \times 20B$. Se vengono usate tutte le 1190 libbre di malto, $35A \times 20B = 1190$. L'insieme dei punti (A, B) che soddisfano questa equazione forma una retta. Tutti i punti (A, B) che corrispondono ai programmi produttivi che richiedono più di 1190 libbre di malto giacciono da una parte della retta e quelli che ne richiedono di meno, dall'altra. Solo quest'ultimo insieme e i punti situati effettivamente sulla retta sono ammissibili a causa delle scorte limitate di malto.

Analogamente, la scarsità di luppolo limita la zona ammissibile a un versante della retta $4A + 4B = 160$, mentre la scarsità di mais limita tale zona a un versante della retta $5A + 15B = 480$. I punti che soddisfano tutte queste condizioni formano la zona ammissibile *(si veda la illustrazione nella pagina precedente)*. Si osservi che la zona ammissibile è convessa: tutti i segmenti di retta che uniscono due dei suoi punti (compresi i punti del perimetro) sono interamente contenuti nella zona stessa.

Dal momento che il profitto del fabbricante di birra è di 13 dollari per ogni barile di ale e di 23 dollari per ogni barile di birra, il suo problema è quello di massimizzare il profitto totale: $13A + 24B$. A tale scopo egli deve trovare un punto (A, B) della zona ammissibile convessa in cui $13A + 23B$ presenti un valore di massimo. Nella programmazione lineare una tale misura dei benefici da massimizzare (o, qualche volta, dei costi da minimizzare) viene chiamata «funzione obiettivo».

La funzione obiettivo può essere incorporata nel grafico della zona ammissibile aggiungendo una terza dimensione. Per ciascun punto (A, B) che rappresenta un programma produttivo, il profitto atteso è dato dall'altezza della funzione $13A + 23B$ al di sopra del piano in quel punto. Il compito che aspetta il fabbricante è ovviamente quello di individuare un punto della zona ammissibile in cui la funzione obiettivo raggiunga la sua massima altezza sul piano. Se si dovesse calcolare tale funzione per tutti i punti, il compito non sarebbe eseguibile; tuttavia le due proprietà distintive del problema, e cioè la convessità della zona ammissibile e la linearità della funzione obiettivo, contribuiscono a facilitare la ricerca.

Poiché la zona è convessa, ogni punto al suo interno può far parte di un segmento di retta i cui estremi giacciono sul contorno della zona (di fatto, per un punto dato si possono tracciare infiniti segmenti di retta; non importa quale viene scelto). Nello spazio che sovrasta ciascuno di questi segmenti di retta è possibile costruire un grafico della funzione obiettivo che dà il profitto di ciascun punto del segmento. Poiché la funzione obiettivo è lineare, il grafico sarà una retta *(si veda l'illustrazione nella pagina a fronte)*. Il grafico della retta può essere parallelo al piano, nel qual caso tutte le strategie produttive situate sul segmento avranno lo stesso profitto. Se il grafico della funzione obiettivo che sovrasta il segmento non è parallelo al piano, deve assumere un valore massimo a uno degli estremi situati sul contorno della zona ammissibile. Di conseguenza, il massimo della funzione obiettivo che sovrasta un tale segmento deve essere raggiunto sempre in uno dei punti di intersezione del segmento con il contorno. Dal momento che la medesima analisi si può applicare a qualsiasi segmento di retta della zona ammissibile, ne segue che il massimo globale della funzione obiettivo si trova invariabilmente in un punto del contorno della zona. Il fabbricante potrà allora, nella sua ricerca del massimo profitto, ignorare tutto l'interno della zona ammissibile e considerare solo quelle strategie produttive che corrispondono ai punti del contorno.

Con un analogo ragionamento è possibile fare un passo avanti nell'analisi. Se la zona ammissibile è un poligono, ciascun punto del contorno giace su un segmento di retta i cui estremi sono due vertici del poligono. Un grafico della funzione obiettivo per tale segmento si può costruire nello stesso modo che per un segmento che interseca la zona. Anche qui il massimo si deve trovare a uno degli estremi (o ad ambedue, se la funzione obiettivo è costante e quindi parallela al piano). In tal modo un valore di massimo della funzione obiettivo relativa all'intera zona ammissibile può essere individuato fra i vertici. Il fabbricante deve solo controllare, al massimo, il valore della funzione su tutti i vertici della zona ammissibile e scegliere quello che garantisce il profitto più alto. Può essere certo allora che nessun'altra strategia produttiva potrebbe assicurare migliori profitti. Nell'esempio considerato qui vi sono cinque vertici. Quello nel punto (12,28), che rappresenta la produzione di 12 barili di ale e di 28 barili di birra, garantisce un profitto di $(12 \times 13) + (28 \times 23)$, ovvero 800 dollari. È questo il massimo profitto conseguibile dal fabbricante.

L'inclusione di un ulteriore vincolo (per esempio una scarsità di lievito) non altererebbe in modo significativo l'interpretazione geometrica del problema del fabbricante di birra. Il poligono potrebbe avere semplicemente sei vertici anziché cinque. D'altra parte l'introduzione di un terzo prodotto modificherebbe il modello più in profondità, rendendolo tridimensionale. Le disuguaglianze in tre variabili corrispondono a semispazi definiti da piani nello spazio bidimensionale. La zona ammissibile non sarebbe più un poligono, ma avrebbe l'aspetto di una gemma tagliata, un politopo tridimensionale le cui facce sono poligoni. Al crescere del numero n delle variabili di decisione, l'interpretazione geometrica conserva la sua validità mentre diventa di più ardua visualizzazione il politopo n-dimensionale definito dalle intersezioni di iperpiani $(n-1)$-dimensionali. I vertici conservano tuttavia la loro funzione specifica e le loro posizioni si determinano mediante metodi algebrici che si sostituiscono all'intuizione geometrica.

Sembrerebbe che, limitando la valutazione della funzione obiettivo ai vertici della zona ammissibile, la soluzione dei problemi di programmazione lineare mediante enumerazione debba diventare più accessibile. Resta tuttavia il fatto che, come nell'applicazione di attribuzione, il numero delle possibilità da enumerare cresce in maniera esplosiva. Un problema con 35 variabili decisionali e 35 vincoli sarebbe del tutto impossibile.

Il metodo del simplesso di Dantzig esamina, sì, i vertici, ma in modo selettivo. Nell'esempio del fabbricante di birra il metodo parte dal vertice situato nell'origine (0,0): dal momento che qui la produzione è nulla, anche il profitto è nullo. Seguendo verso l'esterno l'uno o l'altro degli spigoli incidenti, si arriva a punti con valori di funzione obiettivo sempre maggiori. Il metodo del simplesso sceglie uno di questi spigoli, ad esempio l'asse B, e lo segue fino all'altra estremità, il vertice (0,32); qui il programma produttivo richiede la fabbricazione di 32 barili di birra e niente ale, per un profitto di 736 dollari. A partire da questo vertice uno spigolo incidente conduce a valori ancora superiori della funzione obiettivo. L'algoritmo del simplesso prosegue perciò fino al vertice (12,28) all'altra estremità dello spigolo, dove il profitto di 12 barili di ale e di 28 barili di birra è di 800 dollari. In (12,28) tutti gli spigoli incidenti conducono a posizioni sfavorevoli; l'algoritmo pertanto si arresta con la dichiarazione che il vertice (12,28) è ottimale.

In generale il metodo del simplesso si sposta lungo gli spigoli di un politopo da un vertice all'altro adiacente, continuando a innalzare il valore della funzione

Alla teoria della complessità computazionale è recentemente riuscito di far rientrare la programmazione lineare nell'insieme *P* delle classi di problemi limitati da un polinomio. Questa attribuzione, rappresentata dal punto *A*, è stata resa possibile dalla dimostrazione, proposta dal matematico L. G. Hačan, che un nuovo algoritmo per la programmazione lineare - il cosiddetto metodo dell'ellissoide - è limitato da un polinomio. La funzione polinomio di un numero *n* è una somma finita di potenze di *n*, moltiplicate ciascuna per una costante. Si dice che una classe di problemi è limitata da un polinomio se il numero di operazioni aritmetiche elementari necessarie a risolvere qualsiasi problema della classe è limitato da una funzione polinomio di misura *s* esprimente la dimensione del problema. Prima del risultato di Hačan non si sapeva se per la programmazione lineare esistesse un algoritmo del genere (sebbene non vi fossero prove che non esistesse). Si sapeva che la programmazione lineare rientrava nell'insieme più ampio *NP* delle funzioni polinomiali non deterministiche (punto *B*), dove *NP* è l'insieme di classi di problemi per cui è agevole controllare l'ammissibilità di una soluzione proposta. Un secondo sottoinsieme di *NP* è il cosiddetto insieme dei problemi *NP*-completi; se si dimostrasse che tutte le classi *NP*-complete di problemi sono limitate da polinomi, si avrebbe che tutte le classi di *NP* sono contenute in *P*. In un certo senso, i problemi *NP*-completi sono i più difficili fra quelli contenuti in *NP* (punto *C*).

obiettivo. Il procedimento può partire da qualsiasi vertice e si ferma quando non vi sono più vertici adiacenti con valori della funzione obiettivo maggiori di quelli fino a quel momento raggiunti. Questa regola di arresto è valida solo perché la zona ammissibile è convessa: la convessità garantisce infatti che un vertice di massimo locale sia anche un vertice di massimo globale. Per controllare se vi sia la possibilità di miglioramenti basta guardare solo le immediate vicinanze di ciascun punto.

Come si fa a dire se uno spigolo che parte da un vertice dato migliorerà il valore della funzione obiettivo? La chiave consiste nel concetto di «valore marginale» attribuito a ciascuna risorsa in ogni vertice del politopo ammissibile. Per molti problemi di programmazione lineare pluridimensionali il metodo algebrico comunemente usato per scegliere un percorso da vertice a vertice può essere compreso solo se si comprende il significato dei valori marginali.

Il metodo dell'ellissoide, come quello del simplesso, può essere interpretato geometricamente. Nel caso del problema del fabbricante di birra, esso parte da un grande cerchio E, di centro X e dimensioni definite dai dati del problema, contenente la zona ammissibile. Si costruisce poi una successione di ellissi minori contenenti ciascuna la soluzione ottimale. Dal momento che il centro del cerchio è ammissibile, occorre che ciascuna delle ellissi via via costruite contenga tutti i punti del cerchio la cui funzione obiettivo è almeno altrettanto grande che in X. La più piccola delle ellissi che soddisfano questa condizione è F. Il centro di F, nel punto Y, non è ammissibile. Quindi, l'algoritmo esige che la successiva ellisse G, centrata su Z, contenga tutti i punti di F che giacciono sul versante ammissibile del vincolo $4A + 4B = 160$, già violato dal punto Y. Il procedimento continua tagliando l'ellisse in esame con un contorno della funzione obiettivo se il centro dell'ellisse è ammissibile; se non lo è, l'ellisse successiva circoscrive la parte di quella precedente che soddisfa un vincolo violato dal suo centro. Le aree delle ellissi si contraggono con una rapidità sufficiente ad assicurare la convergenza dei rispettivi centri su una soluzione ottimale.

Nel vertice di massimo del problema del fabbricante di birra vi sono 210 libbre di malto in eccesso, che rimangono inutilizzate nel programma produttivo rappresentato dal vertice suddetto. L'aggiunta o la sottrazione di mezzo chilo di malto dalla disponibilità iniziale di 1190 libbre non modificherebbe il profitto conseguibile. Ma l'aggiunta di un'oncia di luppolo renderebbe possibile un aumento di 2 dollari nel profitto totale. Questo aumento è appunto il valore marginale del luppolo nel vertice di massimo. Lo si può interpretare come il risultato di una «spinta» impressa alla retta del vincolo, che lo fa allontanare dall'origine per riflettere l'ulteriore disponibilità di un'oncia di luppolo *(si veda l'illustrazione nella pagina a fronte)*. Analogamente, il valore marginale del mais in questo vertice è di 1 dollaro.

I prezzi marginali hanno un'interpretazione economica naturale. Se il fabbricante di birra potesse acquistare un'altra oncia di luppolo, il suo profitto aumenterebbe di 2 dollari e, se il luppolo fosse disponibile a un prezzo inferiore ai 2 dollari per oncia, varrebbe la pena acquistarne. D'altra parte, se per tale quantità si pagassero più di 2 dollari, sarebbe più conveniente per il fabbricante distogliere il luppolo dalla produzione e venderlo. Ciò non significa che l'acquisto o la vendita del luppolo debba sempre continuare al prezzo di 2 dollari per oncia, ma per aumenti di circa 19 once o cali fino a 32 once il prezzo di 2 dollari rimane nell'esempio il punto di pareggio economico.

I valori marginali sono qualche volta chiamati anche «prezzi ombra» o «prezzi contabili». Essi indicano di quanto ciascuna risorsa scarsa contribuisce alla convenienza economica di ciascun articolo in produzione. Per esempio, 1 barile di ale richiede 5 libbre di mais con un prezzo contabile di 1 dollaro per libbra, 4 once di luppolo con un prezzo contabile di 2 dollari per oncia e 35 libbre di malto con un prezzo contabile nullo. Il prezzo contabile totale dello ale è pari al profitto di 13 dollari per barile.

Immaginiamo che venga proposta la produzione di un nuovo articolo, la birra leggera. La fabbricazione di 1 barile di birra leggera richiede 2 libbre di mais, 5 once di luppolo e 24 libbre di malto. Quanto profitto per ettolitro si deve ricavare dalla birra leggera in modo da giustificare il dirottamento di risorse dalla produzione di birra normale e di ale? Alla domanda rispondono i prezzi contabili delle rispettive risorse. Il prezzo contabile totale degli ingredienti della birra leggera è $(2\times1) + (5\times2) + (24\times0)$, pari a 12 dollari. Questa cifra misura il profitto che si perderebbe distogliendo risorse dalla fabbricazione della birra normale e dello ale per produrre 1 barile di birra leggera. In tal modo, la produzione di birra leggera sarebbe conveniente se garantisse un profitto di almeno 12 dollari per barile.

La proprietà della convessità nella programmazione lineare stabilisce che per due punti qualsiasi della zona ammissibile un segmento di retta che li congiunga deve giacere dentro tale zona. La zona ammissibile sulla sinistra è convessa; quella a destra è un insieme di punti discreti. La convessità assicura che ogni massimo locale di una funzione obiettivo lineare sia un massimo globale.

La funzione dei prezzi marginali nel giudicare dell'opportunità di introdurre un nuovo prodotto può contribuire a spiegare in che modo il metodo del simplesso individui il percorso sui vertici del politopo capace di aumentare il valore della funzione obiettivo. Supponiamo che il vertice attualmente in esame sia quello in (0,32) nel problema della fabbrica di birra. Innanzitutto si può determinare il prezzo marginale di ciascuna risorsa nel caso di tale programma produttivo. Dal momento che in questo punto il malto e il luppolo sono in eccesso, i loro valori marginali sono nulli. Il mais, invece, scarseggia e di conseguenza il suo valore marginale è positivo. Poiché nel punto (0,32) viene prodotta soltanto la birra normale, il valore marginale degli ingredienti di 1 barile di birra deve essere uguale al profitto legato alla birra, e cioè 23 dollari. Dal momento che solo il mais ha valore marginale diverso da zero e che per 1 barile di birra sono necessarie 15 libbre di mais, il valore marginale del mais è 23 dollari diviso per 15 libbre ovvero circa 1,53 dollari per libbra.

Qual è il valore marginale degli ingredienti dello ale misurato nel vertice (0,32)? Se tale valore è inferiore al profitto ricavabile dalla vendita di 1 barile di ale, il dirottamento di alcune risorse dalla produzione di birra e quella di ale provocherà un aumento nel profitto complessivo del fabbricante. Per fare 1 barile di ale occorrono 5 libbre di mais. Di conseguenza, il valore marginale degli ingredienti di 1 barile di ale è $(23/15) \times 5$, ovvero circa 7,67 dollari per barile. Come nell'esempio della birra leggera, il valore marginale rappresenta la perdita di profitto risultante dal dirottamento delle risorse dalla fabbricazione di birra normale a quella di ale. Dal momento che tale valore è minore di 13 dollari (profitto atteso dalla vendita di ciascun barile di ale), converrebbe fabbricare più ale di quanto non indichi il programma produttivo nel punto (0,32). Di fatto, mantenendo costante il consumo totale di mais e distogliendo il mais dalla birra allo ale, il fabbricante può aumentare il suo profitto per ogni barile di ale di 13 – 7,67 dollari, cioè 5,33

Aumenti marginali nella disponibilità di risorse scarse alterano in modo prevedibile il potenziale profitto di un ipotetico fabbricante di birra. Se si disponesse di un'altra libbra di mais, il profitto massimo crescerebbe di 1 dollaro: tale modificazione della funzione obiettivo si riflette nello spostamento del vertice massimale della zona ammissibile dal punto in colore al circoletto in colore *(grafico in alto)*. Se si disponesse di un'altra oncia di luppolo, il profitto massimale aumenterebbe di 2 dollari *(grafico al centro)*. Una variazione nella disponibilità di malto non altererebbe invece il profitto conseguibile: il malto è già una risorsa sovrabbondante *(grafico in basso)*. Le variazioni del profitto associate alle variazioni di disponibilità di ciascuna risorsa son note con il nome di «prezzi ombra» o «prezzi contabili». La loro funzione è di dirigere l'algoritmo da un vertice all'altro. Le variazioni marginali sono state esagerate per maggiore chiarezza.

	A	B	C
X	6	1	2
Y	7	6	5
Z	6	2	4

I problemi a flusso di rete sono una classe particolare di problemi noti come problemi di programmazione a numeri interi, ai quali si possono applicare le tecniche della programmazione lineare. In un problema di programmazione a numeri interi le variabili di decisione non sono divisibili, ma possono assumere solo valori interi. In generale, questi problemi si trovano nell'insieme degli *NP*-completi e sono perciò considerati difficili. Il costo di ciascun tratto della rete *(diagramma in alto a sinistra)* è dato da una matrice numerica. Si osservi che la matrice è identica a una possibile matrice per un problema di attribuzione 3 × 3.

Le soluzioni dei problemi di attribuzione devono tuttavia essere espresse in valori interi e non pare che i vincoli sul flusso di rete siano sufficienti a garantire tale risultato. Sono molti i flussi frazionari che rispondono ai vincoli *(diagramma in alto a destra)*. Ciononostante, le soluzioni estremali di problemi a flusso di rete trovate con metodi di programmazione lineare sono sempre espresse in valori non frazionari, purché siano interi i valori dei flussi in arrivo e in partenza su ciascun nodo. Il problema dell'attribuzione può essere interpretato come un caso speciale del problema a flusso di rete *(diagramma in basso)*.

dollari. Essendo tenuto costante il quantitativo di mais, una maggiore produzione di ale corrisponde a uno spostamento lungo lo spigolo della zona ammissibile che rappresenta il vincolo sulle scorte di mais. È esattamente questo il modo in cui il metodo del simplesso riconosce la possibilità di migliorare la funzione obiettivo lungo uno spigolo. Qui lo spigolo collega il vertice (0,32) al vertice (12,28) del poligono.

Anche in problemi di programmazione lineare piuttosto modesti la zona ammissibile può presentare un numero enorme di vertici e vi è sempre la possibilità che il metodo del simplesso richieda un numero enorme di iterazioni, cioè di mosse da vertice a vertice, per risolvere il problema. Di solito, tuttavia, la ricerca selettiva considera solo una frazione insignificante del numero totale di vertici. Coloro che risolvono regolarmente problemi con 2000÷3000 vincoli sulle risorse a 10 000÷15 000 variabili decisionali constatano che l'operazione richiede solitamente da pochi minuti a poche ore di funzionamento di un potente elaboratore. Vorremmo sapere qualcosa di più preciso sull'efficienza del metodo del simplesso.

Un sistema per convincere uno scettico dell'efficienza di un algoritmo è quello di dare una certa garanzia sotto forma di una funzione G e di un enunciato che affermi che per un problema di dimensione d il tempo di esecuzione dell'algoritmo non supererà $G(d)$. La dimensione d del problema è una misura della quantità di informazione (espressa sovente in termini di numero delle cifre binarie) necessaria a specificare i dati in entrata. È questa la cosiddetta «analisi dell'eventualità peggiore», dove il valore della garanzia è definito dal problema di dimensione d più complesso e difficile che l'algoritmo può essere chiamato a risolvere. Il tempo di esecuzione dell'algoritmo per la soluzione di un problema di dimensione d dipende dalle caratteristiche del particolare elaboratore impiegato. Per rendere indipendente dal tipo di macchina usato la misura $G(d)$, la funzione è solitamente espressa non in ore o minuti bensì in termini del massimo numero di operazioni aritmetiche elementari (addizione, moltiplicazione e confronto) che devono essere eseguite dall'algoritmo.

Se $G(d)$ è opportunamente piccola per i valori di d che ci si aspetta di incontrare, si sarà inclini a riconoscere l'efficienza dell'algoritmo proposto. Così, per esempio, Harold W. Kuhn della Princeton University ha messo a punto un algoritmo per l'applicazione di attribuzione che richiede al massimo cn^3 operazioni elementari per risolvere un problema n per n, dove c è una costante piccola. Quando n è pari a 2, 3, 4 o 5, l'algoritmo può effettivamente richiedere più tempo dell'enumerazione (per quanto con un elaboratore nessuno dei due metodi richiederebbe più di un batter di ciglio). Una volta che n cresca di valore, $n!$ (che misura la dimensione dell'enumerazione) supera molto rapidamente cn^3. In realtà $n!$ cresce in maniera talmente esplosiva che nessun elaboratore attualmente esistente sarebbe in grado di portare a termine i calcoli. D'altra parte, n^3 cresce con minore rapidità al crescere di n. Per $n = 35$ l'algoritmo di Kuhn può ancora risolvere il problema in un batter d'occhio. Inoltre, al crescere di n da 35 a 36, il peso dei calcoli per l'algoritmo di Kuhn cresce di un fattore di circa $(36/35)^3$, ovvero approssimativamente 1,09. Il lavoro richiesto invece dall'enumerazione crescerebbe di un fattore 36.

Coloro che operano nel settore degli elaboratori sono particolarmente interessati agli algoritmi che vanno sotto il nome di «algoritmi limitati da un polinomio», nei quali cioè la funzione G è una funzione polinomiale di d. Una funzione polinomiale di d è una somma di più termini, in ciascuno dei quali d è elevato a una potenza costante e moltiplicato per un coefficiente costante. Se d è sufficientemente grande, qualsiasi algoritmo la cui garanzia dell'eventualità peggiore cresca come $d!$ o come una funzione esponenziale di d, ad esempio 2^d, procederà con maggiore lentezza di uno limitato da un polinomio. Di conseguenza, coloro che lavorano con gli elaboratori considerano la proprietà di essere limitato da un polinomio un criterio teorico per valutare l'efficienza di un algoritmo.

Dal momento che il metodo del simplesso funziona molto bene nella pratica, si è tentati di pensare che si tratti di un algoritmo limitato da un polinomio. Di fatto, si è trovato nella pratica che una buona regola empirica approssimativa è che il numero delle iterazioni mediante criterio del simplesso cresce in funzione lineare del numero dei vincoli. Negli impieghi abituali, tuttavia, il metodo del simplesso non è limitato da polinomi. Victor La Rue Klee, Jr., della Università di Washington, e George J. Minty, Jr., della Indiana University, hanno costruito una famiglia infinita di problemi di programmazione lineare in cui il metodo del simplesso è sostanzialmente inadeguato al pari dell'enumerazione.

Che un algoritmo così valido nella pratica possa essere considerato tutt'altro che efficiente in teoria può apparire strano. In realtà non bisogna dimenticare che il criterio teorico esamina i risultati dell'eventualità peggiore. Nella pratica non sembra che, almeno fino a oggi, si presentino esempi come quelli costruiti da Klee e Minty che provocano un cattivo funzionamento del metodo del simplesso. Per spiegare la divergenza fra l'esperienza pratica con il metodo del simplesso e la sua prestazione nell'eventualità peggiore, i ricercatori hanno tentato di dimostrare che in un certo senso probabilistico gli esempi cattivi sono davvero «patologici». Tuttavia, nonostante gli interessanti risultati in questa direzione riferiti recentemente da Dantzig, la dimostrazione si scontra con difficoltà considerevoli.

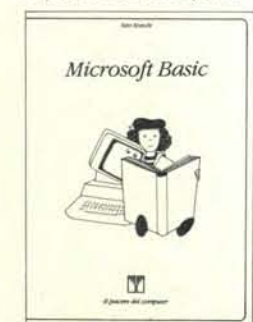
Vi è una certa flessibilità nel modo in cui il metodo del simplesso sceglie a ogni iterazione fra più vertici adiacenti, quando ogni vertice sembrerebbe migliorare il valore della funzione obiettivo. La costruzione di Klee-Minty mostra che una strategia molto diffusa conduce a risultati non polinomiali. Altri ricercatori hanno a tutt'oggi costruito esempi analoghi per altre strategie. Non si sa tuttavia se esista una realizzazione del metodo del simplesso che sia limitata da un polinomio. Fino a poco tempo fa non si sapeva ancora se esistesse un algoritmo limitato da un polinomio per la programmazione lineare; era un problema irrisolto di grande interesse teorico. Infine L. G. Hačan dell'Accademia delle scienze dell'URSS ha mostrato che un algoritmo messo a punto da tre altri matematici sovietici, N. Z. Šor, D. B. Judin e A. S. Nemirovskij, poteva essere realizzato in tempo polinomiale.

Dal momento che il lavoro di Šor, Judin e Nemirovskij non pone condizioni di linearità, il nuovo algoritmo differisce notevolmente dal metodo del simplesso. Si tratta del cosiddetto metodo dell'ellissoide. Il punto di partenza è dato da un grande ellissoide centrato nell'origine. Più in generale, l'ellissoide di partenza è una sfera con un raggio abbastanza grande da contenere sicuramente la soluzione di ottimo. L'algoritmo passa a costruire progressivamente ellissoidi sempre più piccoli, ciascuno dei quali è tale da contenere la soluzione di ottimo. Il volume degli ellissoidi che via via si succedono si contrae abbastanza in fretta da garantire la convergenza dei rispettivi centri sulla soluzione di ottimo. Non è necessario, per usare l'algoritmo, disegnare gli ellissoidi; semplici formule ricorsive bastano a descriverli. Se si considera il lavoro necessario a costruire ciascun ellissoide a partire da quello che l'ha preceduto, il limite al numero complessivo delle operazioni è dell'ordine di $(m+n)n^6l$, dove m è il numero dei vincoli, n quello delle variabili decisionali e l il logaritmo in base 2 del coefficiente più grande che figura nei dati del problema.

Tale espressione è un limite polinomiale, ma, come qualsiasi altro limite di questo tipo, può crescere ancora di molto per valori grandi di m, n e l. È possibile che il metodo dell'ellissoide, come quello del simplesso, dia migliori risultati nella pratica di quanto non lasci pensare il suo limite peggiore? I risultati preliminari non sono incoraggianti. Le prove di calcolo hanno finora mostrato una convergenza estremamente lenta per il metodo dell'ellissoide. Vi sono inoltre delle famiglie assai semplici di problemi che possono far funzionare il metodo sui tempi lunghi previsti dal limite. Ciò contrasta con la forma complessa dei problemi costruiti per sconfiggere il metodo del simplesso.

Un'altra difficoltà rilevante del metodo dell'ellissoide riguarda le condizioni di memoria. I problemi di programmazione lineare molto grandi tendono nella pratica a presentare una struttura diradata: la proporzione di coefficienti diversi da zero nei dati è di solito piuttosto piccola. È possibile utilizzare il metodo del simplesso in modo da conservare questa struttura rada mentre i dati vengono elaborati nel corso dello svolgimento dell'algoritmo. Si tratta di una proprietà decisiva, dal momento che basta immagazzinare nell'elaborazione i soli coefficienti non nulli. Invece, non sembra possibile sfruttare nel metodo dell'ellissoide questa proprietà.

Sebbene deboli appaiano fino a questo momento le prospettive di un impiego pratico del metodo dell'ellissoide, bisogna ancora vedere quali saranno i suoi sviluppi. Nel breve tempo trascorso dopo la nota di Hačan molto lavoro è stato fatto, con alcuni miglioramenti sostanziali. Tuttavia le difficoltà qui sottolineate appaiono davvero molto temibili.

La programmazione lineare è uno strumento pratico di grande interesse teorico. Per il matematico rappresenta una teoria delle disequazioni lineari, e cioè un elegante contrappunto alla teoria classica delle equazioni lineari. Per le amministrazioni pubbliche e private costituisce un valido aiuto nella gestione a livello decisionale e per la programmazione a lungo raggio. Probabilmente la conseguenza più importante dell'exploit teorico di Hačan sarà l'aumento di interesse per gli aspetti pratici e per quelli teorici della programmazione lineare.

Piombo e argento nell'antico Egeo

Durante l'età del bronzo questi due metalli venivano estratti dagli stessi minerali. Lo studio delle abbondanze tipiche degli isotopi del piombo ha dimostrato che i giacimenti sfruttati erano in pratica solamente due

di Noël H. Gale e Zofia Stos-Gale

Nel Mediterraneo orientale l'Età del bronzo, e in particolare i suoi ultimi secoli, furono un periodo di splendore senza precedenti. Da Troia sulla costa asiatica sino all'Egitto e oltre, ai grandi palazzi di Creta e di Micene, sulla terraferma greca, leghe a base di rame trasformarono le arti della guerra e della pace e consentirono di accumulare enormi tesori, costituiti soprattutto da gioielli, vasellame e altri oggetti d'argento e d'oro.

La documentazione archeologica ha sollevato vari interrogativi. Quale fu la natura delle antiche tecniche estrattive e metallurgiche che resero disponibili questi metalli? Quale funzione svolse nell'antica metallurgia il piombo? Le necessarie tecniche metallurgiche si svilupparono autonomamente nell'Egeo oppure si erano diffuse in questa regione in conseguenza di progressi realizzati in precedenza nel Vicino Oriente? E infine, da quali giacimenti venivano estratti i metalli? Per esempio, è stato spesso asserito che Creta e Micene importavano rame da Cipro, oro dall'Egitto e argento dall'Asia Minore. Queste asserzioni sono vere?

La natura dei contatti culturali e commerciali fra i vari popoli è un argomento di grande interesse nella preistoria dell'Egeo, ma è stato solo scarsamente chiarito dalle limitate scoperte di ceramiche appartenenti a una cultura nell'area geografica di un'altra o da somiglianze stilistiche spesso assai dibattute per esempio, fra pugnali od oggetti d'oro e d'argento. La possibilità di determinare il luogo d'origine del metallo di cui è fatto un oggetto, unitamente alla scoperta e alla datazione di antiche miniere e dei resti di antiche installazioni metallurgiche, ci offrono oggi l'opportunità di dare risposte molto più precise a tali domande.

Studi condotti all'Università di Oxford nel corso degli ultimi cinque anni hanno permesso di identificare, al di là di ogni ragionevole dubbio, le località di provenienza di molti oggetti d'argento e di piombo dell'Età del bronzo nell'Egeo. Qui daremo un rendiconto del lavoro che ha messo a disposizione degli studiosi della preistoria dell'Egeo i risultati forniti da una varietà di discipline: analisi chimica e fisica, analisi degli isotopi, geologia, metallurgia e tecnologia mineraria.

L'Età dei metalli, che ebbe inizio nel mondo mediterraneo in epoca imprecisata attorno alla metà del IV millennio a.C., vide grandi mutamenti nella natura delle società arcaiche. Non pare lecito dubitare del fatto che lo sviluppo della capacità di estrarre metalli per fusione dai loro minerali fu un fattore importante ai fini del passaggio dalla società relativamente indifferenziata ed essenzialmente agricola del tardo Neolitico alla società, fondata sulla specializzazione artigianale e sempre più strutturata in diversi livelli gerarchici, che emerse nell'Egeo con la Fase I dell'Età del bronzo antica designata EB I (da *Early Bronze Age*, dal 3500 a.C. circa al 2900 circa) e si sviluppò sempre più rapidamente nella Fase II (EB II, dal 2900 circa al 2100 a.C. circa).

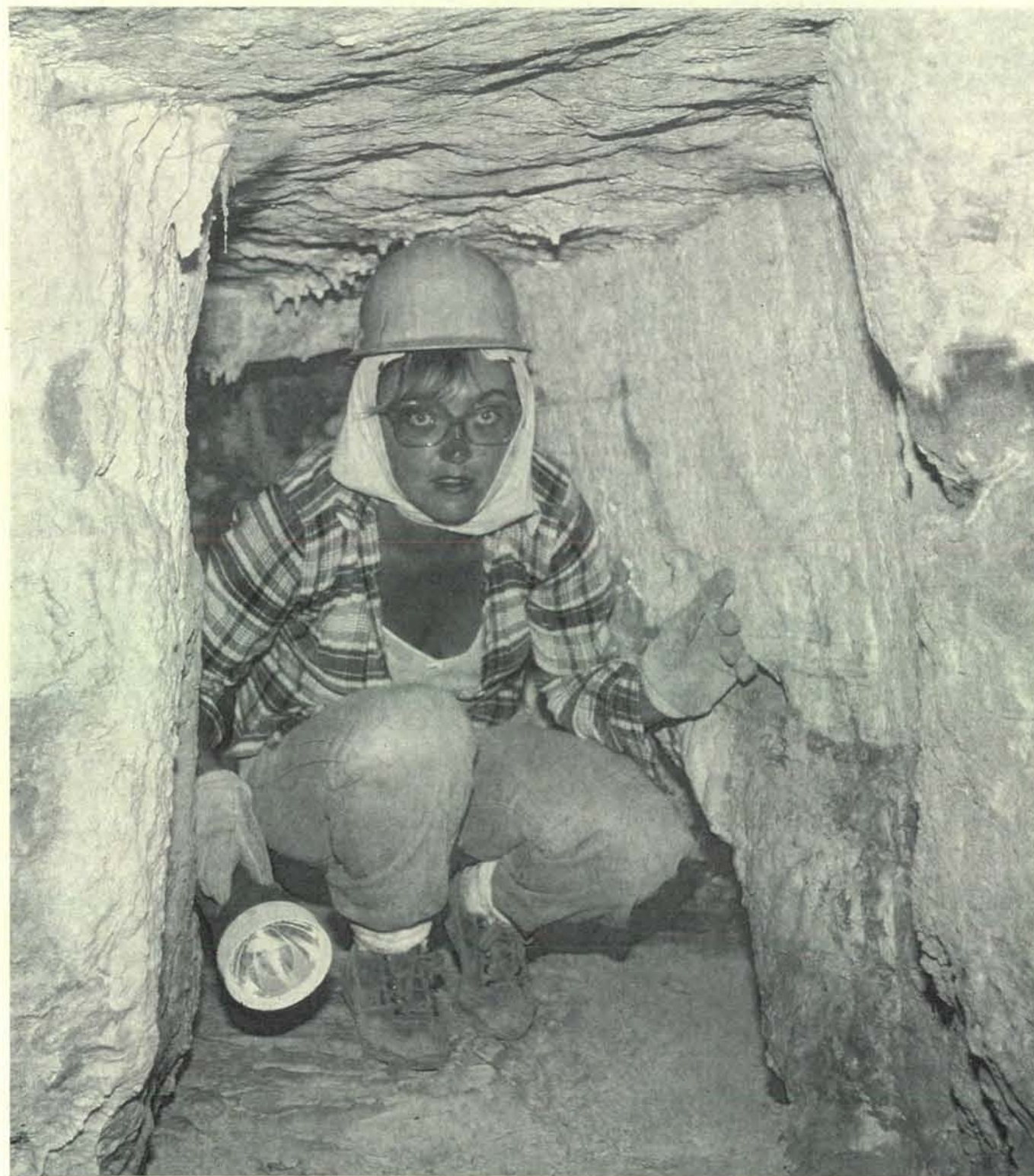
Gli studiosi della preistoria della metallurgia e del suo impatto sulla società sono stati inclini a concentrare la loro attenzione sul rame e sulle sue leghe, il bronzo arsenicale e stannico, per due ragioni. La prima ragione è la supposta importanza di questi metalli utilitari ai fini del mutamento sociale, adatti com'erano alla produzione di armi di qualità superiore e allo sviluppo di specializzazioni artigianali attraverso l'introduzione di nuovi utensili. La seconda ragione deriva dal presupposto che la documentazione archeologica confermi la tesi che il primo metallo ottenuto per fusione dai suoi minerali sia stato il rame. Entrambe queste ragioni possono oggi essere contestate.

In primo luogo l'impatto di una metallurgia di qualsiasi genere sulla società dipende dal riconoscimento di qualità desiderabili nel metallo prodotto. Tali qualità sono di due tipi. Una è l'utilità, una qualità che i vari tipi di bronzo senza dubbio posseggono. L'altra potrebbe essere definita estetica: il riconoscimento che certi metalli hanno un aspetto attraente e sono rari. Queste sono le qualità possedute tanto dall'oro quanto dall'argento. Benché nessuno dei due metalli fosse particolarmente utile a fini puramente pratici, entrambi furono ben presto molto apprezzati e gli oggetti fatti d'oro e d'argento divennero un possesso ambito. In effetti si potrebbe sostenere che l'emergere di concetti di ricchezza, di potere e di status gerarchico associati al possesso d'oro e d'argento contribuirono a mutare la natura delle società arcaiche non meno della diffusa utilizzazione di leghe del rame.

La convinzione che il rame sia stato il primo metallo ottenuto per fusione dai suoi minerali può essere contestata per le ragioni seguenti. È ben noto che quantità abbastanza grandi di rame si trovano allo stato di metallo nativo ed è molto probabile che questa forma non combinata del rame sia stata quella utilizzata per prima dall'uomo. Non pare però che il rame nativo sia stato molto apprezzato nella produzione di armi o di utensili o che sia stato sfruttato in grandi quantità, se si può giudicare dal numero relativamente scarso di esemplari sicuri che troviamo nella documentazione archeologica. Se l'uomo avesse utilizzato il rame nativo, è difficile pensare che questa risorsa potesse già essere esaurita nell'Età del bronzo antica al punto da rendere necessario il passaggio alla fusione dei minerali.

L'incentivo a far ricorso a questa tecnica di estrazione non sarebbe stato inoltre probabilmente molto forte. Utensili e armi di rame nativo avrebbero dovuto competere in un primo tempo con quelli di selce, di ossidiana e di altri tipi di pietra e non avrebbero goduto al confronto di un grande vantaggio. Nell'Egeo l'intervallo iniziale di un millennio fra la prima comparsa di oggetti in rame e la rivoluzione metallurgica della Fase II dell'Età del bronzo antica può essere attribuito probabilmente al ritardo nello sviluppo della capacità di produrre leghe di bronzo arsenicale e stannico, con le loro superiori proprietà metallurgiche.

Il bronzo arsenicale fu scoperto quasi certamente per caso quando furono fusi minerali di rame che contenevano impurezze arsenicali. Il bronzo naturale così prodotto dovette rivelarsi presto molto più duro anche del rame nativo martellato. A questa scoperta dovette seguire abbastanza presto la produzione intenzionale di bronzo arsenicale con minerali contenenti arsenico.

La fusione del rame è piuttosto difficile. A meno che i primi artigiani del metallo non avessero già acquisito esperienza nella fusione, più facile, di minerali ridotti, difficilmente avrebbero potuto tentare tale processo con minerali del rame come la malachite, un minerale di colore verde scuro, la azzurrite, di colore azzurro, o anche la calcopirite, di un color giallo ottone (che assomiglia almeno in parte al rame nativo per colore, ma che dev'essere arrostita prima di poter essere

Una galleria di una miniera nella regione del Laurio, a sud-est di Atene. La galleria veniva scalpellata con precisione nella roccia calcarea, aprendo uno spazio alto circa un metro e largo mezzo in cui si poteva procedere solo carponi. Quando la galleria raggiungeva una vena o una massa più consistente di galena, il minerale contenente piombo e argento, la galleria veniva allargata e il minerale veniva asportato. Quella illustrata in questo caso è una galleria nella miniera numero 5 a Thorikos. Essa fu aperta probabilmente nel V secolo a.C. L'estrazione di minerali nel Laurio ebbe inizio migliaia di anni prima, nell'Età del bronzo. Nella fotografia è riconoscibile la coautrice, Zofia Stos-Gale.

fusa). Benché la malachite e l'azzurrite possano essere ridotte a metallo a temperature molto inferiori al punto di fusione del rame (1083 °C), il rame rimane disseminato e non disponibile finché la temperatura non sale abbastanza da fonderlo e da trasformare la ganga, costituita dai minerali rocciosi non desiderati, allo stato di scoria fluida; il risultato di questo processo è la presenza di due liquidi non miscibili sul fondo della fornace.

In pratica la fusione di tutt'e tre questi minerali comuni del rame richiede una temperatura di circa 1200 °C. Questo fatto è sufficiente a confutare la vecchia ipotesi fantasiosa secondo cui la fusione del rame sarebbe stata scoperta quando un fuoco da campo fu acceso casualmente su un affioramento di un minerale di rame: i fuochi da campo non raggiungono infatti una tale temperatura. Invece il più comune fra i minerali del piombo, la galena (sulfuro di piombo), si fonde così facilmente che in effetti sarebbe possibile estrarne piombo con un fuoco da campo circondato da un cerchio di pietre. Esperimenti eseguiti da Ronald Tylecote, dell'Istituto di archeologia dell'Università di Londra, hanno dimostrato che il piombo può essere ottenuto per fusione dalla galena in un fuoco che bruci carbone di legna o legna secca a temperature inferiori a 800 °C. (Questa temperatura è molto superiore al punto di fusione del piombo metallico, che è di 327 °C.) È quindi molto più probabile che sia stata casuale la scoperta dell'estrazione per fusione del piombo dalla galena, minerale caratterizzato da una lucentezza metallica.

Quali sono i documenti più antichi sulla fusione del piombo? A differenza del rame nativo, il piombo nativo è raro, cosicché quando in un sito archeologico si trova del piombo, esso è stato ottenuto quasi certamente mediante la fusione di un minerale. Nel sito urbano di Çatal Hüyük, in Asia Minore, sono state rinvenute perline di piombo che risalgono al 6500 a.C. circa. Piombo è stato trovato anche in un contesto del VI millennio a.C. a Yarim Tepe nell'Iraq, nel sito del V millennio di Arpachiyeh nell'Iraq e nei siti del IV millennio di Anau e di Hissar III in Iran e di Naqada in Egitto. Questi reperti suggeriscono che la fusione del piombo, forse su piccola scala, abbia avuto inizio già nel VII millennio a.C.

L'identificazione di rame ottenuto sicuramente per fusione è più difficile. Poiché il rame nativo non è raro, il ritrovamento di manufatti di rame in contesti archeologici molto antichi non è necessariamente una prova dell'estrazione del rame per fusione dai suoi minerali. I manufatti potrebbero essere stati prodotti martellando semplicemente rame nativo. Neppure le analisi chimiche consentono di distinguere un manufatto di rame nativo abbastanza puro da uno prodotto con rame ottenuto per fusione da un minerale abbastanza puro. Inoltre i minerali del rame come la malachite o l'azzurrite trovati in siti archeologici antichi potrebbero essere stati usati come pigmenti e come cosmetici e non per l'estrazione del rame mediante fusione.

L'unica documentazione certa dell'estrazione del rame per fusione è quella fornita da reperti rappresentati da manufatti di rame arsenicale o di rame puro associati a scorie di fusione certe. Pochissimi fra i manufatti di rame più antichi riescono a superare entrambi questi esami. I reperti che superano tutte le prove suggeriscono che l'antichità del rame ottenuto per fusione è sostanzialmente minore di quella del piombo estratto con lo stesso procedimento. Gli esempi più antichi di rame fuso provengono dai livelli dell'inizio del IV millennio a Tepe Yahya in Iran. Le miniere di rame a Rudna Glava in Yugoslavia sono datate al 3700 a.C. e rame fuso si trova nei siti della metà del IV millennio di Sitagroi III e di Kephala in Grecia e di Amuq in Siria. Si può sostenere perciò che il primo metallo a essere estratto per fusione dall'uomo sia stato il piombo, con un margine di vantaggio sul rame che, sulla base della documentazione attualmente disponibile, si avvicinerebbe a 3000 anni.

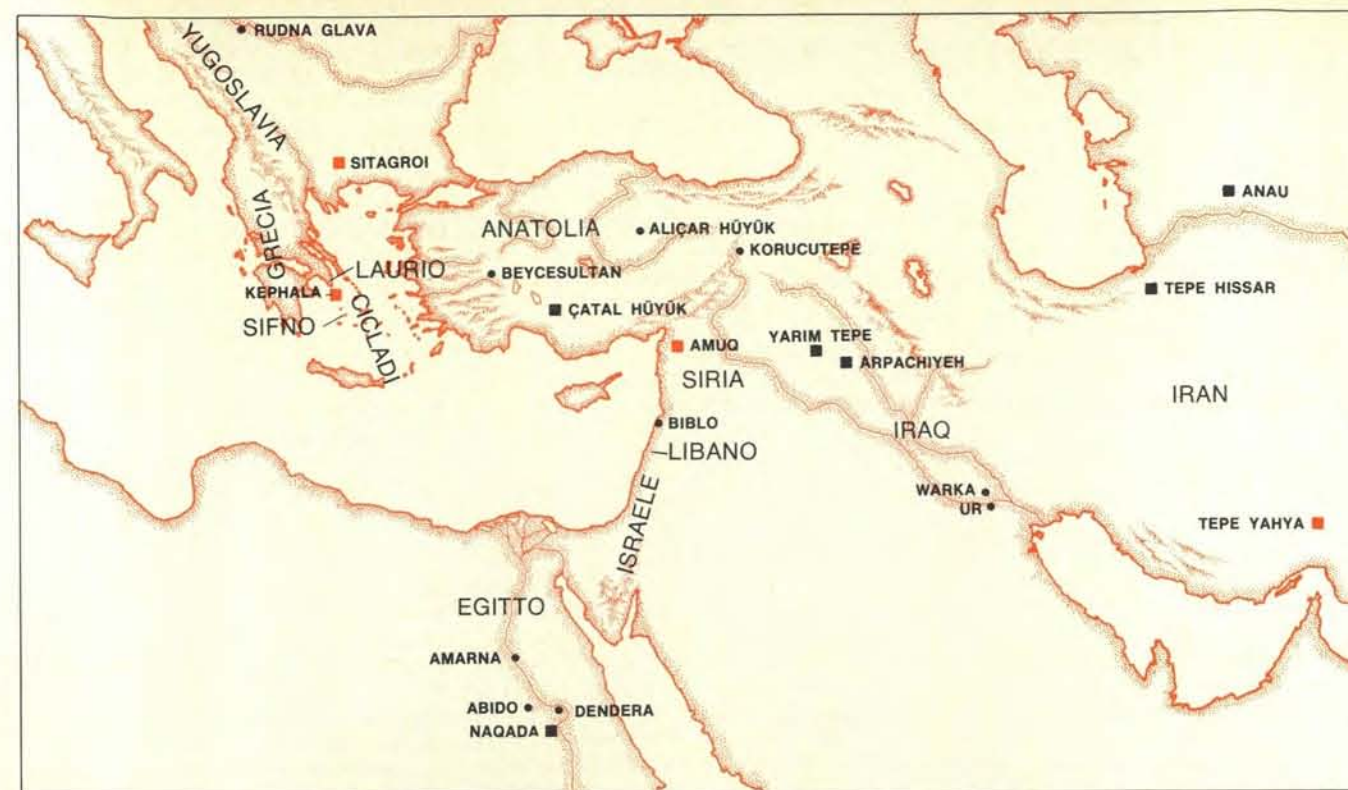
Che cos'ha a che fare tutto questo con l'oro e con l'argento? Per concentrare la nostra attenzione sull'argento, questo metallo allo stato nativo non è raro come il piombo, ma la sua abbondanza è pari solo allo 0,2 per cento rispetto a quella del rame nativo. Esso è perciò troppo scarso per poter essere stato una fonte importante di questo metallo nell'antichità. L'argento è presente però in minerali del piombo come la galena e in vari minerali complessi di piombo-antimonio-argento. Inoltre, quando tali minerali vengono fusi, l'argento si liquefa assieme al piombo, mentre altri elementi presenti nel minerale, come ferro, manganese, silicio, calcio e alluminio, passano principalmente nelle scorie.

L'argento viene separato dal piombo per mezzo del procedimento noto come coppellazione. La lega di piombo e argento viene fusa in un crogiolo e mantenuta a una temperatura di circa 1100 °C mentre su di essa viene soffiata aria. L'aria ossida il piombo, trasformandolo in litargirio (monossido di piombo). Le impurità, come rame, stagno, antimonio, arsenico e bismuto, vengono in gran parte ossidate assieme al piombo; l'argento (contenente una traccia d'oro) non viene ossidato e, una volta che il litargirio sia stato assorbito dalle pareti del crogiolo (o eliminato con mezzi meccanici), rimane come residuo un globulo fuso d'argento. L'argento ottenuto per coppellazione contiene sempre una quantità residua di piombo, che può variare dal 2 allo 0,05 per cento.

La tecnica della coppellazione del piombo per estrarne l'argento potrebbe essere stata scoperta nella preistoria della metallurgia contemporaneamente alla tecnica della fusione del rame, e forse ancor prima. In una necropoli della metà del IV millennio a Biblo, nel Libano, sono stati rinvenuti più di 200 manufatti di argento. Nessuno di essi è stato però finora analizzato per determinare se l'argento di cui sono fatti sia stato o no ottenuto per coppellazione. Lo stesso si può dire per i pochi oggetti di argento della fine del IV millennio che sono stati rinvenuti in Palestina, a Ur e a Warka in Mesopotamia e a Beycesultan, Aliçar Hüyük e Korucutepe in Asia Minore. In Egitto sono stati rinvenuti almeno 26 manufatti di argento di età predinastica, ossia anteriore al 3000 a.C. Noi ne abbiamo analizzato uno, un coperchio di uno scrigno d'argento proveniente da Naqada, eseguito attorno al 3600 a.C. Esso ha un contenuto in piombo dello 0,4 per cento ed è perciò sicuramente un esempio di argento ottenuto per coppellazione.

La scoperta che l'argento poteva essere ottenuto sottoponendo a coppellazione il piombo ricco d'argento deve avere stimolato un crescente interesse per la fusione del piombo. Il piombo in sé è un metallo tenero e malleabile, diventa facilmente opaco e non fu dapprima di molta utilità. In effetti potrebbe non essere casuale il fatto che la comparsa di grandi quantità di manufatti d'argento nel IV millennio coincida o preceda di poco le prime indicazioni dell'inizio dell'estrazione del rame mediante fusione dei suoi minerali. È lecito infatti attendersi che un'accresciuta attività di fusione di un gruppo di minerali possa condurre a un aumento dell'efficienza delle pratiche di fusione in generale.

La forma più antica di fusione del piombo va probabilmente identificata con la fusione della galena, un procedimento abbastanza facile dato che la galena fonde a basse temperature. Lo stimolo fornito dalla richiesta di argento, però, dovette rendere necessaria la fusione di minerali contenenti molte impurità, in cui la galena era mescolata con ganga. La separazione del piombo da tali minerali deve aver richiesto una temperatura di fusione di circa 1200 °C e anche l'aggiunta sperimentale di altri minerali come fondenti per facilitare la separazione della ganga sotto forma di scoria fusa. Esperimenti simili con altri minerali di aspetto promettente e con temperature di fusione più elevate potrebbero avere preannunciato l'Età del bronzo.

Il fatto che l'argento ottenuto per coppellazione contenga piccole quantità di piombo consente, per così dire, di prendere le impronte digitali degli antichi manufatti d'argento. Le analisi chimiche comuni degli oggetti metallici possono rivelare qualcosa sul modo in cui il metallo fu estratto dal minerale o consentire di stabilire la prima comparsa del bronzo arsenicale o stannico, ma nel caso del piombo, dell'argento e di manufatti del rame non legati, tali analisi hanno un'utilità limitata nel determinare il tipo del minerale o la localizzazione della miniera. La ragione di ciò va vista nel fatto che i processi di fusione e di coppellazione - con l'aggiunta di fondenti, la separazione degli elementi fra le scorie e il metallo e le differenze della rapidità con cui i vari metalli volatilizzano - distruggono la distribuzione degli elementi principali, secondari e di quelli presenti solo in tracce caratteristica del minerale originario.

Tracce di fusione di minerali all'inizio dell'Età dei metalli sono state trovate nei siti qui indicati. I quadratini neri rappresentano la fusione di piombo e i quadratini in colore quella di rame. La documentazione più antica della fusione di minerali di rame è costituita dai reperti di Tepe Yahya nell'Iran. Le tracce più antiche di fusione di minerali di piombo sono state rinvenute, invece, a Çatal Hüyük, in Turchia.

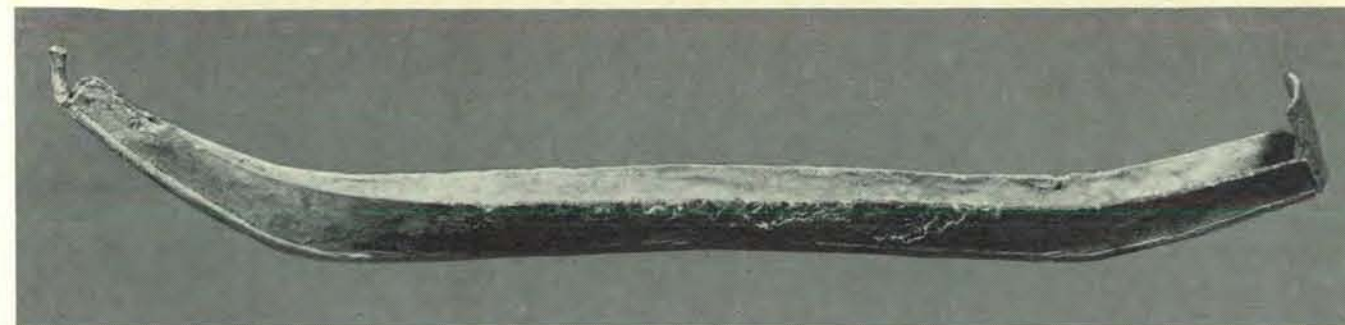
In questa cartina sono indicati giacimenti di minerali, molti dei quali non sono stati sfruttati nell'Età del bronzo o nel periodo della Grecia classica. I quadratini neri designano i minerali di piombo, quelli in colore i minerali contenenti oro. I triangoli neri pieni indicano antiche miniere di piombo e i triangoli neri vuoti indicano siti in cui sono state trovate tracce di attività di fusione, come mucchi di scorie e litargirio (monossido di piombo). I triangoli pieni con pallino colorato in centro rappresentano siti dove sono in corso attività minerarie moderne.

L'isola di Sifno nelle Cicladi fu famosa in epoca classica come luogo di estrazione di argento. Oggi si sa che i minerali di Sifno contenevano un miscuglio di piombo, antimonio e argento. Essi venivano scavati in un certo numero di siti sulle coste e nell'interno. Sia la datazione con il carbonio 14 sia la datazione mediante il metodo della termoluminescenza dimostrano che i minerali ad Agios Sostis, sulla costa nord-orientale, venivano estratti già nell'Età del bronzo antica. Essi venivano scavati per estrarne principalmente l'argento, ma si utilizzava anche il piombo.

La penisola del Laurio, con le miniere di piombo (*quadratini*) era la fonte più famosa di argento in epoca classica. Lo studio della composizione isotopica di manufatti antichi e i reperti nella miniera n. 3 a Thorikos dimostrano che i minerali venivano estratti già nell'Età del bronzo antica.

L'analisi isotopica fornisce invece dati precisi sulla proporzione dei quattro isotopi del piombo (Pb-204, Pb-206, Pb-207 e Pb-208) presenti nel metallo. I processi chimici di fusione, raffinamento e corrosione lasciano inalterata la composizione isotopica originaria; ciò consente di comparare direttamente le proporzioni degli isotopi del piombo presenti in un manufatto con quelle dei minerali provenienti da varie miniere dell'antichità. La composizione in isotopi del piombo di diversi minerali del piombo o di piombo e argento differisce per il fatto che una parte degli atomi di piombo 206, di piombo 207 e di piombo 208 sono il prodotto del decadimento radioattivo rispettivamente dell'uranio 238, dell'uranio 235 e del torio 232. Minerali diversi avranno una diversa composizione in isotopi del piombo a seconda della loro età geologica e delle quantità relative di uranio e di torio presenti in origine nei liquidi che formarono il minerale.

Gli autori di questo articolo hanno speso mezzo decennio in uno studio generale della protometallurgia del piombo, dell'argento e dell'oro nel mondo egeo (compreso l'Egitto). Le prime fasi della nostra ricerca furono intraprese in collaborazione con ricercatori dell'Istituto Max Planck per la fisica nucleare a Heidelberg, i quali ci hanno aiutato nel lavoro sul campo, nella datazione e in alcune analisi chimiche. Il nostro studio si proponeva vari obiettivi. Intendevamo trovare, studiare e datare varie miniere antiche dell'Egeo, analizzare le tecniche minerarie dell'epoca e identificare e studiare resti di forni, scorie, litargirio e altri residui della fusione e della coppellazione. Per mezzo di analisi chimiche e isotopiche del piombo di antichi manufatti e di minerali provenienti dalle varie miniere, ci attendevamo inoltre di scoprire le fonti mineralogiche e geografiche del piombo, dell'argento e dell'oro a disposizione delle varie culture dell'Egeo dall'Età del bronzo sino all'epoca greca classica. Queste ricerche avrebbero dovuto fornirci a loro volta informazioni su contatti culturali e itinerari commerciali, argomenti importanti sia per lo studioso della preistoria sia per lo storico del periodo più antico delle attività estrattive e metallurgiche. Nell'area egea esistono numerosi depositi d'oro e di piombo ricco d'argento. Molti sono menzionati nella letteratura classica, per esempio dal drammaturgo Eschilo (525-456 a.C.), dallo storico Erodoto (490-425 a.C.) e dal geografo Strabone (63 a.C. - 21 d.C.). I giacimenti più famosi sono quelli, ricchi di piombo e argento, della zona del Laurio, in prossimità dell'estremità della penisola dell'Attica, una quarantina di chilometri a sud-est di Atene. Nel V secolo a.C. l'argento del Laurio era la fonte principale della potenza della città-stato di Atene.

Altre miniere menzionate da Erodoto, che noi abbiamo trovato ed esplorato, sono le miniere d'oro sulla costa sudorientale dell'isola di Taso (sulla costa occidentale della quale sono molte le antiche miniere di piombo argentifero), alcune fra le antiche miniere della regione del Pangeo in Macedonia (sulla terraferma a nord di Taso) e le miniere di piombo argentifero dell'isola di Sifno nelle Cicladi (90 chilometri a sud-est della regione del Laurio). Abbiamo raccolto anche campioni di minerali da miniere moderne e da giacimenti di minerali non sfruttati attorno all'Egeo e abbiamo sottoposto ad analisi campioni di scorie e di litargirio.

Delle 31 regioni di mineralizzazione del piombo e dell'argento da noi studiate, molte non furono probabilmente sfruttate nell'Età del bronzo e alcune neppure in tempi classici. Anche nei casi in cui abbiamo trovato tracce di attività minerarie «antiche» o antichi cumuli di scorie, per lo più non è stato ancora determinato il periodo di attività ed è possibile che questa si situi in epoca posteriore all'Età del bronzo. I tentativi di datare una miniera sulla base dei tipi di utensili rinvenuti in essa o, in assenza di utensili, del tipo di segni lasciati dagli utensili sulle pareti delle gallerie, sono dubbi in quanto gli sviluppi della tecnica nell'antica mineralogia furono estremamente lenti. Una datazione attendibile richiede la lunga e faticosa ricerca, nelle gallerie di miniere, di carbone databile con la tecnica del carbonio 14, o di cocci di ceramica databili sulla base di considerazioni stilistiche o di un'analisi della termoluminescenza. La presenza di resti di carbone si deve all'antico metodo di rompere la roccia accendendo a contatto di essa un fuoco e poi gettandovi contro dell'acqua o forse anche al metodo di aerare una miniera accendendo un fuoco alla base di un pozzo verticale per creare un tiraggio ascendente.

Le importanti miniere del Laurio contengono galena abbastanza ricca d'argento (fra 500 e 5000 grammi di argento per tonnellata). Le tracce dello sfruttamento delle miniere del Laurio per opera degli ateniesi e di altri greci nel V e IV secolo a.C. sono state molto studiate da geologi e archeologi a partire dall'Ottocento. La regione è un alveare di antiche miniere dalle quali venivano estratti minerali che si trovavano per lo più nelle zone di contatto fra scisto e marmo. I minerali venivano in parte separati dalla ganga già all'interno delle miniere e poi erano trasportati fuori. Qui venivano rotti in pezzi più piccoli e ulteriormente separati su tavole di lavaggio rettangolari o elicoidali di forma ingegnosa. L'acqua per il lavaggio veniva accumulata in grandi cisterne rivestite con un cemento idraulico impermeabile fatto col litargirio ottenuto come prodotto secondario della coppellazione. Pare che uno fra i centri più importanti della zona del Laurio sia stato quello di Thorikos, che è stato scavato da Herman Mussche e Paule Spitaels dell'Università di Gand. Il sito comprendeva un tempio, un teatro del VI secolo a.C., edifici pubblici, officine metallurgiche, attrezzature per il lavaggio del minerale, miniere e un porto.

Erodoto scrive che le miniere di Sifno avevano fatto di quest'isola la più ricca di tutte le Cicladi, così ricca che una decima di argento fu sufficiente a costruire un tesoro in marmo riccamente decorato nel santuario di Apollo a Delfi. Pausania ricorda la leggenda che in seguito le miniere di Sifno furono inondate e il flusso di argento dall'isola terminò a causa dell'ira di Apollo quando gli abitanti di Sifno smisero di pagare le decime dovute.

Una zona di mineralizzazione di ferro e manganese attraversa Sifno da nord-nordest verso sud-sudovest nella breccia marmorea del Terziario. Le antiche miniere si trovano proprio all'interno di questa zona di mineralizzazione del ferro e manganese. L'estrazione del minerale di ferro nell'Ottocento inferse danni considerevoli alle gallerie scavate dagli antichi minatori, i quali ignoravano il ferro, ma erano così efficienti nello sfruttare i depositi di piombo argentifero che oggi riesce difficile trovarvi traccia del minerale di piombo.

Segni lasciati da utensili sulle pareti delle miniere sono abbondanti in queste gallerie strette e irregolari. Qua e là nelle pareti sono scavate nicchie per sistemarvi lampade a olio. Un carattere tipico della tecnologia mineraria a Sifno (osservato però anche nelle miniere di Taso) è che la maggior parte delle rocce di scarto, asportate per estendere le gallerie; non sono state portate in superficie ma accumulate sul posto sotto forma di pareti formate da materiale di scarto.

Le miniere nella parte settentrionale dell'isola, sulla penisola di Agios Sostis, si trovano al livello del mare, e alcune gallerie scavate al di sotto del livello del mare sono oggi inondate, in accordo con la leggenda riferita da Pausania. Antiche gallerie si trovano anche più in alto nell'interno dell'isola, ad Agios Sylvestros, Vorini, Kapsalos e Xero Xylon. Tracce di antiche attività minerarie sono state trovate anche ad Agios Joannis, il sito di una miniera moderna, e scorie e litargirio sono stati rinvenuti a Plati Yialos e Kapsalos.

Nelle prime due stagioni del nostro lavoro sul campo a Sifno non trovammo campioni di minerali antichi. Infine ne trovammo piccole quantità nella maggior parte dei siti di Sifno da noi studiati. Si trattava di minerali di piombo-antimonio-argento altamente ossidati: giallastri, pesanti, friabili e di grana fine. Le analisi degli isotopi del piombo di questi minerali e di campioni di scorie e di litargirio ci hanno consentito di definire il «campo» caratteristico degli isotopi del piombo nei minerali di Sifno che è diverso da quello di qualsiasi altra isola delle Cicladi e differisce anche dal campo dei minerali del Laurio.

Press'a poco nello stesso periodo, seguendo un suggerimento di Colin Renfrew, che è ora all'Università di Cambridge, il nostro gruppo di Oxford estese i suoi studi risalendo oltre il periodo classico e cominciò a cercare le miniere di piombo e d'argento sfruttate dalle culture dell'Età del bronzo nella regione dell'Egeo. In principio esaminammo quattro modellini di navi in piombo che erano stati trovati in una tomba del periodo Antico-cicladico (3400-2100 a.C.) sull'isola di Nasso e che erano stati poi donati all'Ashmolean Museum di Oxford da un noto studioso di archeologia minoica, sir Arthur Evans. Con nostra sorpresa, trovammo che il campo isotopico dei modellini di navi in piombo, come pure il loro alto contenuto in antimonio, suggerivano come luogo d'origine del metallo l'isola di Sifno. La conclusione era inevitabile: l'estrazione di minerali a Sifno doveva essere cominciata già nell'Età del bronzo antica.

Un'indagine sistematica delle miniere di Agios Sostis a Sifno ha ora confermato questa conclusione. L'intero versante orientale della penisola è traforato da un labirinto di antiche gallerie di miniere. Le gallerie non hanno un tracciato regolare; esse si addentrano nelle zone del minerale

Il modellino di nave in piombo appartiene a un gruppo trovato in una tomba dell'Età del bronzo antica nell'isola di Nasso. Oggi è all'Ashmolean Museum di Oxford. L'analisi isotopica del piombo indica che il modellino è formato da piombo estratto a Sifno all'inizio del III millennio a.C.

di ferro e manganese alla ricerca solo delle sacche relativamente scarse del minerale di piombo, antimonio e argento.

Nel corso dei millenni le pietre che costituiscono i muri di materiale di scarto all'interno delle gallerie si sono cementati naturalmente in conseguenza dell'infiltrazione di liquidi calcarei. Uno di tali muri è stato smantellato con cura allo scopo di determinare la posizione stratigrafica (e quindi l'età relativa) di qualsiasi materiale estraneo in esso rinvenuto. La parte superiore del muro, meno saldamente cementata, conteneva un frammento di ceramica ricoperto da una vernice nera. La determinazione dell'età del frammento col metodo della termoluminescenza ha fornito due date che ci riportano entrambe a quello che gli storici definiscono il periodo arcaico: 600±100 a.C. e 530±140 a.C. Nella parte inferiore, più compatta del muro, furono trovati altri due cocci di ceramica più grossolana assieme a resti di carbone. Le età, determinate col metodo della termoluminescenza, erano rispettivamente 2590 ± 440 a.C. e 2780 ± 400 a.C. Le età determinate col carbonio 14 dei campioni di carbone sono state 2970±180 a.C. e 2610±50 a.C.

Questa prova del fatto che le miniere di Agios Sostis furono sfruttate già nell'Età del bronzo antica è stata confermata anche da altre scoperte fatte nella stessa area. I percussori di pietra trovati nelle gallerie sono di un tipo caratteristico dell'Età del bronzo antica. Alle scorie e al litargirio disseminati alla superficie sono mescolate in quantità schegge di ossidiana, una delle quali ha la forma di una punta di freccia con codolo perfettamente conformata; esse potrebbero essere tutte manufatti dell'Età del bronzo antica. Infine, fra i frammenti di ceramica ritrovati in superficie, c'è un coccio che presenta una decorazione impressa opaca, tipica della ceramica del tardo Neolitico e dell'Età del bronzo antica nella regione, e un altro che reca impresso il disegno a lisca di pesce tipico del periodo noto come Antico-cicladico I (3400-2900 a.C.).

I popoli dell'Egeo dell'Età del bronzo impiegavano il piombo per vari usi. Il metallo è stato trovato, in contesti neolitici tardi, anche sulle isole di Makronesos e di Zea, appena al largo della costa dell'Attica, di fronte alla regione del Laurio. Il piombo compare nei reperti dell'Età del bronzo antica su otto isole delle Cicladi (Antiparo, Dhespotikon, Zea, Sira, Amorgo, Io, Milo e, come abbiamo già visto, Nasso). Esso è noto anche in contesti dell'Età del bronzo antica in insediamenti cicladici sulla costa dell'Attica (Raphin e Agios Kosmas), a Creta, a Lesbo, a Lemno e a Troia.

Il piombo veniva utilizzato nella produzione di dischi per fusi, piombi per reti da pesca e filo metallico (spesso trovato nelle tombe) e serviva anche per riparare ceramiche rotte. Era usato a scopi decorativi per anelli e braccialetti e per costruire figurine umane e modelli di navi. Nella tarda Età del bronzo il metallo fu utilizzato anche per pesi da bilancia, che avevano come unità un peso di circa 61 grammi. Questi pesi sono stati trovati in tutto l'Egeo in contesti dell'Età del bronzo media e tarda. Un numero cospicuo di questi pesi è stato riportato in luce ad Akrotiri sull'isola di Santorino (Tera), nei pressi di Creta, e nell'isola di Zea.

Nell'Età del bronzo l'argento veniva usato nella produzione di ciotole e di tazze, ma soprattutto era impiegato per gioielli: diademi, collane, braccialetti, anelli e spille per abiti. I pezzi più elaborati, fra cui alcuni in piombo, venivano sepolti con i loro proprietari, assieme ad altre suppellettili funerarie, come ceramiche, armi e utensili in bronzo. Quest'usanza, che toglieva regolarmente dalla circolazione tanti oggetti così apprezzati, deve aver sostenuto una domanda di argento che avrebbe altrimenti potuto essere soddisfatta, almeno in parte, dalla trasmissione in eredità. (Dal punto di vista della moderna ricerca le pratiche funerarie dell'Età del bronzo rappresentano una vera fortuna; esse rendono infatti poco probabile la fusione di gioielli per consentirne la riutilizzazione assieme ad altri oggetti d'argento di origine diversa, con la distruzione delle indicazioni sulla loro origine contenute nelle proporzioni originarie degli isotopi del piombo.)

Un confronto degli oggetti rinvenuti in varie tombe cicladiche ci fornisce informazioni sull'emergere di una stratificazione sociale già nell'Età del bronzo antica. Dapprima una tale stratificazione si manifesta solo sotto forma di piccole variazioni nella quantità e qualità delle suppellettili funerarie. Successivamente si riflette in una concentrazione di oggetti preziosi in tombe che si distinguono per l'assenza dei tipi più comuni di ceramica e che talvolta contengono oggetti principeschi come diademi e ciotole d'argento. Questa tendenza sarebbe culminata più avanti, nel corso dell'Età del bronzo, con gli splendidi tesori in oro e argento nelle tombe a fossa a Micene e con le ricchezze contenute anche in tombe meno importanti altrove, come la tomba a *tholos* (in forma di alveare) a Vapheio nei pressi di Sparta, in Laconia.

Fino a oggi la provenienza del piombo e dell'argento usati dalle culture dell'Età del bronzo dell'Egeo è stata del tutto incerta, anche se molti archeologi hanno suggerito che tali metalli dovevano provenire dall'Asia Minore o dalle lontane regioni occidentali della penisola iberica. Keith Branigan ha suggerito che il piombo e l'argento dovevano essere stati estratti da giacimenti nell'Egeo stesso, ma che tali giacimenti potrebbero essersi esauriti nel corso della media Età del bronzo; nella tarda Età del bronzo tali fonti dovettero quindi essere sostituite da altre fonti, nel Mediterraneo occidentale, nell'Asia Minore o forse nell'Europa centrale. Noi cominciammo a investigare questo problema prendendo l'avvio dalla documentazione archeologica dell'Antico-cicladico.

Fra gli oggetti d'argento noti appartenenti a tale periodo vi sono due diademi, almeno 13 braccialetti, sette piatti, due collane e almeno cinque spille d'argento. Fra gli oggetti in piombo vi sono tre figurine, i quattro modellini di navi di Nasso, tre pesi, un sigillo, vari chiodi per la riparazione di ceramiche e vari frammenti. Questi materiali sono stati rinvenuti principalmente sulle isole di Amorgo, Nasso, Sira, Antiparo, Dhespotikon, Io, Makronesos e Policandro. Per la maggior parte sono stati trovati in tombe assieme a ceramica, figurine di marmo dalle braccia piegate tipiche delle Cicladi, ciotole in marmo, oggetti in bronzo e altre suppellettili funerarie. Questi manufatti ci forniscono una documentazione preziosa della vitalità e dell'originalità della cultura dell'Età del bronzo antica delle Cicladi, i cui prodotti rappresentavano uno fra i punti più alti dell'arte preistorica.

Questo braccialetto d'argento dell'Età del bronzo antica, trovato in una tomba nell'isola di Amorgo, nelle Cicladi, è uno dei 30 manufatti d'argento rappresentativi del periodo. L'analisi isotopica di sette di tali manufatti indica che uno di essi è composto da metallo proveniente da Sifno e cinque da metallo del Laurio. Il braccialetto di Amorgo è uno di quelli del Laurio.

Grazie alla generosità dei curatori di musei di Oxford, Londra, Liverpool e Atene abbiamo avuto la possibilità di analizzare 16 di tali oggetti, ossia circa un terzo del totale. Il quadro che ne emerge è sorprendentemente semplice. Sei dei 16 oggetti sono fatti di metallo proveniente dalle miniere del Laurio e otto di metallo proveniente dalle miniere di Sifno. I due restanti sono fatti di metallo proveniente da un giacimento non ancora identificato con sicurezza. Abbiamo trovato giacimenti di minerali di piombo aurifero in almeno altre sei isole delle Cicladi oltre che su Sifno; su due di tali isole, Sira e Serfanto, non solo i minerali sono più ricchi in argento, ma ci sono anche segni di antiche attività minerarie.

DATA a.C.	CRETA E AREE VICINE	CICLADI	GRECIA CONTINENTALE	ASIA MINORE	EGITTO	NOTE
NEOLITICO FINALE 4000-3500						
EB I 3500-2900						
				TROIA I		2970 ± 180: DATAZIONE AL CARBONIO 14 (SIFNO); circa 2800: NAVI DI PIOMBO DI NASSO
EB (ETÀ DEL BRONZO ANTICA)	ANTICO-MINOICO I-II-III	ANTICO-CICLADICO I-II-III	ANTICO-ELLADICO I-II-III			2610 ± 50: DATAZIONE AL CARBONIO 14 (SIFNO)
EB II 2900-2100				TROIA II-V		
					X-XII DINASTIE (2175-1991)	
ETÀ DEL BRONZO MEDIA 2100-1600	MEDIO-MINOICO I-II-III	MEDIO-CICLADICO	MEDIO-ELLADICO	TROIA VI	XII-XIII DINASTIE (1991-1668)	1930: PRIMI PALAZZI CRETESI; 1700: DISTRUZIONE DEI PALAZZI; 1550: TOMBE A FOSSA (MICENE)
			TARDO-ELLADICO I-II			
ETÀ DEL BRONZO TARDA 1600-1100	TARDO-MINOICO I-II-III		MICENEO III (A-C)	TROIA VII	XVIII DINASTIA (1570-1293)	1400: CADUTA DI CNOSSO; 1200: CENTRI MICENEI DISTRUTTI

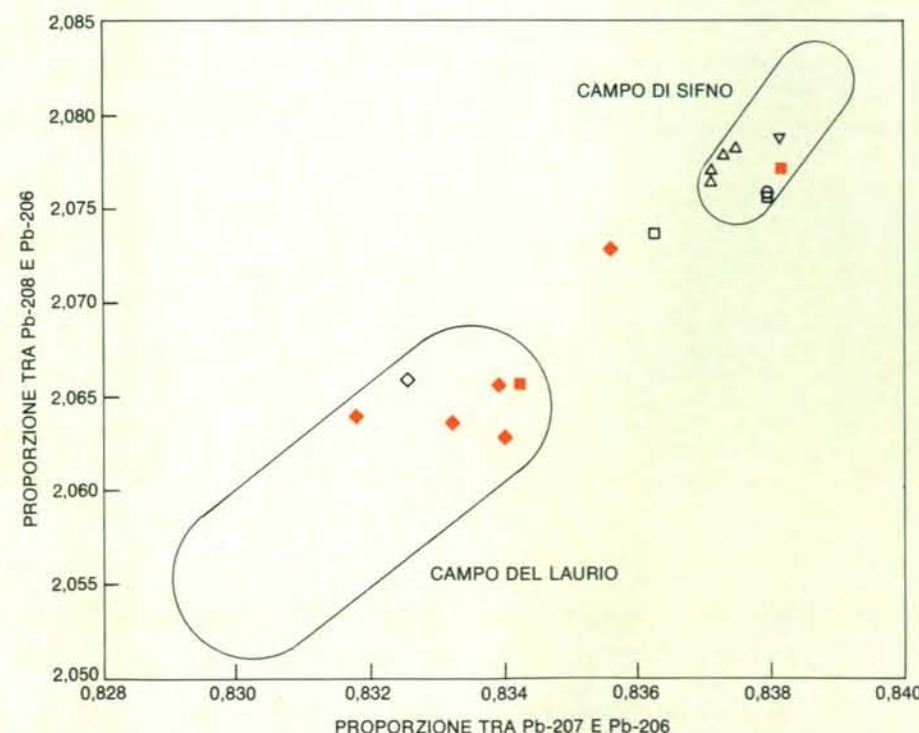
Cronologia dell'Età del bronzo nell'Egeo: tutte le date sono approssimative. La tabella include anche periodi comparabili in Asia Minore e in Egitto, e riporta gli sviluppi fondamentali dell'Egeo. Le attività di estrazione del piombo nell'Egeo potrebbero essere anteriori al 3000 a.C.

La situazione delle fonti del metallo si mantenne immutata anche più avanti nel corso dell'Età del bronzo? In un tentativo preliminare di rispondere a questa domanda, abbiamo analizzato 24 manufatti di piombo venuti in luce ad Akrotiri sull'isola di Santorino. Akrotiri, la Pompei dell'Età del bronzo, era un esteso insediamento quando fu sepolta sotto la cenere dell'esplosione vulcanica che squarciò Tera attorno al 1500 a.C. Le abitazioni signorili, gli affreschi, gli edifici, il tracciato delle strade, gli impianti igienici e le condutture, la ceramica e altri manufatti conservatisi nelle ceneri suggeriscono che la società di Tera doveva essere ricca, molto civilizzata, dipendente dal commercio marittimo e caratterizzata da un apprezzamento vivace e originale delle arti.

La situazione di Akrotiri rispetto alla Creta minoica è incerta. Non si sa se Akrotiri si sia sviluppata indipendentemente da Creta, pur godendo di un fiorente commercio con essa e restando politicamente autonoma, se sia stata conquistata e governata direttamente da un qualche centro di palazzo cretese o se i cretesi vi abbiano stabilito una colonia autonoma. La presenza di una grande quantità di piombo, di un po' d'argento e di litargirio suggerisce che a Santorino siano stati prodotti sia piombo sia argento. Le nostre analisi chimiche e isotopiche di questo materiale attestano la medesima dicotomia quanto alla provenienza che abbiamo trovato negli oggetti rinvenuti altrove nelle Cicladi, ma con un notevole spostamento nelle proporzioni. Soltanto uno degli oggetti in piombo rinvenuti ad Akrotiri è di metallo estratto a Sifno; gli altri 23 sono di metallo proveniente dalle miniere del Laurio.

Che cosa si può dire di Creta? Nell'antica Età del bronzo nelle Cicladi i manufatti di piombo e d'argento costituiscono il 40 per cento circa di tutti i reperti di oggetti in metallo e l'oro è estremamente raro. A Creta si riscontra una situazione inversa. Fra tutti i reperti in oggetti di metallo pubblicati datati all'Età del bronzo antica, soltanto il 3 per cento è costituito da piombo e argento, mentre l'oro è relativamente comune. I ritrovamenti di oggetti di piombo e d'argento sono concentrati nel nord dell'isola, fra Candia (Herakleion) e Mochlos, un'area in cui gli oggetti d'oro e di rame sono invece più rari. Quest'area si trova direttamente dirimpetto alle Cicladi, e i suoi manufatti dell'Età del bronzo antica documentano chiaramente un contatto con la cultura cicladica. Sembra perciò ragionevole il suggerimento di Branigan che gli abitanti della parte settentrionale di Creta possano avere avuto piombo e argento dalle Cicladi. Non abbiamo ancora potuto analizzare questi oggetti, ma speriamo di poter presto estendere il nostro lavoro sui materiali cretesi includendovi manufatti dell'Età del bronzo antica.

Finora abbiamo esaminato i manufatti di piombo e d'argento rinvenuti in soli due siti cretesi, Cnosso e Kommos, e solo per il periodo compreso fra il Medio-minoico e il Tardo-minoico (circa 2150-

Le analisi isotopiche riportano in ascissa la proporzione fra il piombo 207 e il piombo 206 e in ordinata quella fra il piombo 208 e il piombo 206 per determinare il rapporto fra i minerali di piombo delle varie miniere dell'Egeo. Tutti i campioni di minerale provenienti da Sifno *(in colore)* rientrano nel campo a destra in alto. Tutti quelli del Laurio *(in nero)* sono compresi all'interno del campo a sinistra in basso. I campioni da Nanfio *(pallini in colore)*, Serfanto *(rombi in colore)* e Termia *(quadrato in colore vuoto)* si dispongono alla periferia della distribuzione del Laurio, mentre nessuno dei campioni provenienti da altre quattro isole delle Cicladi o dalla miniera di Kallianou nell'Eubea si avvicinano significativamente al campo di Sifno o a quello del Laurio.

I rapporti fra gli isotopi del piombo in 16 manufatti dell'Antico-cicladico, nove di piombo *(in nero)* e sette d'argento *(in colore)*, ne situano 14 all'interno del campo di Sifno o di quello del Laurio. I quattro modellini di navi in piombo rinvenuti a Nasso *(triangoli vuoti)*, una statuina di piombo da Antiparo *(pallino)*, un frammento di piombo da Cheiromylos *(triangolo capovolto)*, un peso in piombo *(quadratino)* e un diadema d'argento da Sira sono all'interno del campo di Sifno, mentre quattro braccialetti d'argento da Amorgo *(rombi pieni)*, una spilla d'argento da Sira *(quadratini)* e un oggetto di piombo da Amorgo *(rombo vuoto)* sono all'interno del campo del Laurio. Una ciotola d'argento da Amorgo e un chiodo di piombo da Sira sono incerti, essendo formati forse da un miscuglio di metalli del Laurio e di Sifno o da metallo estratto vicino a Troia.

1300 a.C.). Nell'età del bronzo media, Creta sviluppò una ricca civiltà molto articolata gerarchicamente, caratterizzata da grandi palazzi, città fiorenti, un brillante artigianato e ampi collegamenti oltremare. La distruzione dei primi palazzi a Cnosso e a Festo alla fine dell'Età del bronzo media (circa 1700 a.C.) non indica in alcun modo un declino della cultura minoica. La documentazione archeologica non conserva alcuna traccia di nuovi venuti, e la ricostruzione dei palazzi annunciò il massimo episodio nella civiltà minoica: il secondo periodo palaziale (1700-1450 a.C.). Gli artigiani dell'isola produssero in questo periodo i loro lavori più belli: la ceramica, le gemme, l'oreficeria e gli affreschi di quest'epoca presentano un livello di sensibilità estetica e di abilità tecnica che assicura loro un posto indiscusso nella storia dell'arte. In questo periodo si sviluppò anche la scrittura minoica del lineare A.

Esiste una documentazione abbondante dei commerci minoici (o di altri tipi di contatti) fuori di Creta nel secondo periodo palaziale. C'erano traffici col Vicino Oriente, con l'Egitto e, su distanze meno lunghe, con le Cicladi e con l'Argolide, sulla terraferma greca. Perciò il piombo e l'argento, che praticamente non esistono a Creta, potrebbero essere venuti da uno qualsiasi di questi luoghi.

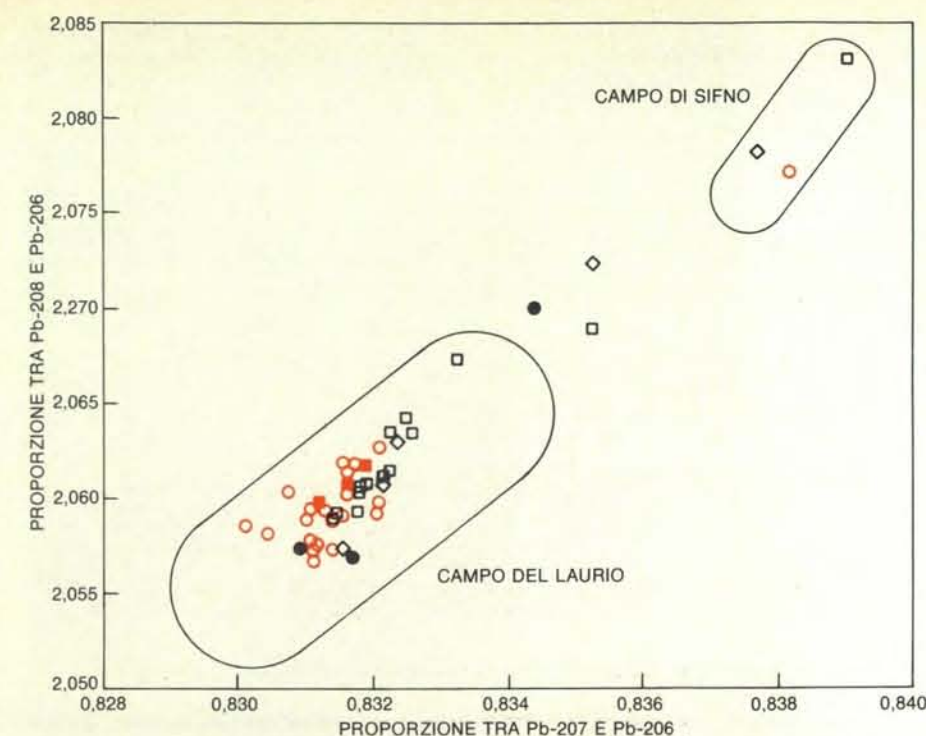
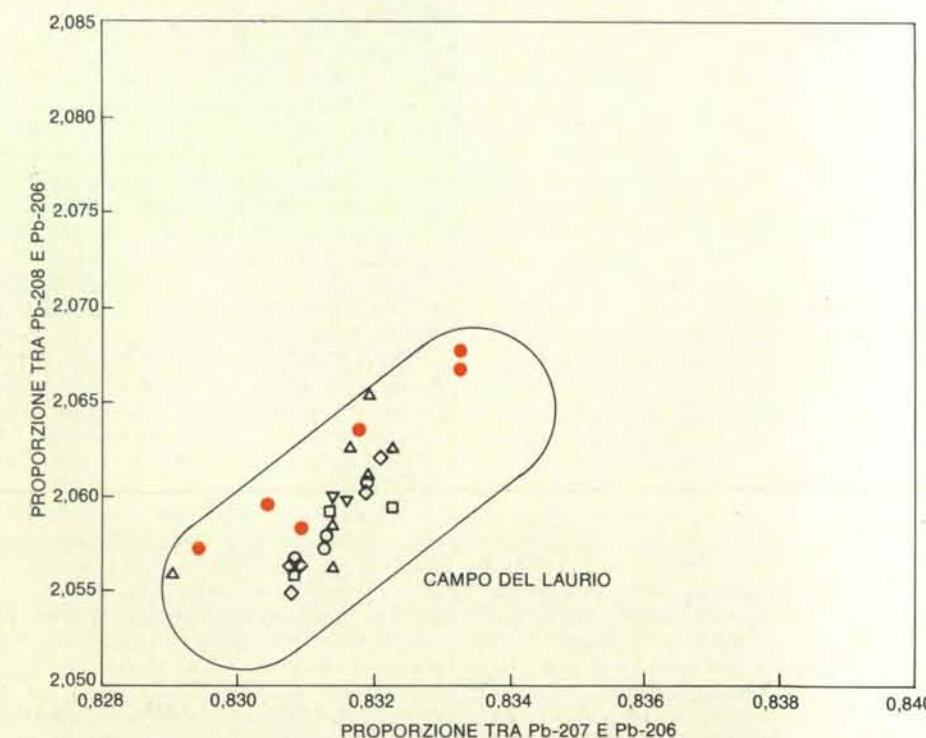
Attorno al 1450 a.C. quasi tutti i siti principali su Creta furono distrutti; la civiltà minoica non riuscì mai più a tornare allo splendore originario. Cnosso e le località circostanti furono occupate dai micenei fra il 1400 circa e il 1340 a.C. Se i micenei avevano allora accesso alle miniere del Laurio, ci si potrebbe attendere di trovare piombo e argento del Laurio almeno nei livelli di Cnosso datati a questo periodo (Tardo-minoico III A).

Il nostro lavoro preliminare dimostra in effetti che la fonte principale dei manufatti in piombo rinvenuti a Cnosso e risalenti a tutto il periodo compreso fra il Medio-minoico e l'occupazione micenea fu il Laurio. L'80 per cento dei manufatti da noi analizzati sono di piombo del Laurio. Solo il 10 per cento sono di piombo di Sifno, benché quest'isola sia molto più vicina a Creta. Il restante 10 per cento provengono da una fonte che non è stata ancora identificata. A Kommos la scarsità dei reperti suggerisce che il piombo fosse una merce di difficile reperimento. Noi abbiamo analizzato gli unici tre manufatti in piombo dell'Età del bronzo scavati in tale sito: due sono di piombo del Laurio, il terzo è di piombo proveniente da una fonte non identificata. Il generale dominio del Laurio come luogo d'origine del piombo usato a Creta nell'Età del bronzo media e recente ribadisce la situazione che abbiamo già trovato a Santorino. Il numero esiguo dei reperti di piombo proveniente da giacimenti non identificati suggerisce però che Creta avesse un accesso limitato a luoghi di produzione del piombo diversi dal Laurio e da Sifno.

Le nostre analisi di manufatti di piombo e d'argento rinvenuti nelle Cicladi, a Santorino e a Creta ci forniscono chiare indicazioni del fatto che le miniere del Laurio furono sfruttate nell'Età del bronzo almeno a partire dal 2900 a.C. (Antico-cicladico II) e sino al 1350 a.C. circa (Tardo-minoico III A). Un'ulteriore conferma dell'attività mineraria nel Laurio in questo intervallo di tempo potrebbe essere fornita da analisi di manufatti di piombo e d'argento rinvenuti in siti dell'Età del bronzo nell'Attica stessa. Nella speranza di poter trovare conferma, abbiamo analizzato vasellame e filo di piombo rinvenuti nei livelli micenei dal 1400 al 1150 a.C. (dal Miceneo III A al III C) nell'Acropoli e nell'Agorà di Atene, piombo e litargirio rinvenuti in un contesto anteriore a Thorikos (1700-1550 a.C., ultima parte del Medio-elladico) e infine manufatti di piombo e di argento provenienti dalla vicina Perati (1190-1100 a.C.). Tutti questi manufatti rivelarono la medesima composizione isotopica (e le medesime caratteristiche chimiche) del piombo estratto dalle miniere del Laurio.

Una prova del fatto che la produzione delle miniere del Laurio raggiungeva località relativamente lontane nella terraferma greca è fornita dall'analisi di quattro pesi per bilancia rinvenuti in Laconia. I pesi, di cui abbiamo dimostrato la composizione con piombo del Laurio, furono trovati nella stessa tomba a *tholos* a Vapheio che conteneva tazze d'oro decorate in rilievo con immagini di tori cretesi.

Fra il 1870 e il 1880 Heinrich Schliemann trovò un favoloso tesoro nelle famose tombe a fossa di Micene risalenti al 1550 a.C. circa (Tardo-elladico I). Di solito l'attenzione si concentra sugli oggetti d'oro, ma le tombe contengono anche grandi quantità d'argento e anche alcuni oggetti in piombo. Qual era l'origine del tesoro, e quale spiegazione si può dare dell'ascesa di Micene al tempo delle tombe a fossa?

Sir Arthur Evans pensava che il carattere minoico di molti oggetti contenuti nelle tombe attestasse l'occupazione di Micene da parte di una dinastia cretese conquistatrice. Quest'opinione fu confutata però da un contemporaneo di Evans, lo studioso austriaco Georg Karo, e oggi il lavoro di Ellen Davis, della City University of New York indica che la maggior parte degli oggetti d'oro rinvenuti nelle tombe a fossa è di origine micenea. Lo stesso Karo cercò una spiegazione dell'improvvisa acquisizione di ricchezze da parte di Micene e del carattere dei tesori nell'ipotesi che molti fra gli oggetti rinvenuti nelle tombe fossero stati saccheggiati a Creta nel corso di incursioni da parte di Micene; i micenei non avrebbero potuto acquistare le loro ricchezze col commercio perché, secondo Karo, non avevano risorse naturali proprie. Il nostro lavoro su questo problema non è molto avanzato, ma noi abbiamo nondimeno dimostrato che al tempo delle tombe a fossa il piombo del Laurio giungeva a Micene. Schliemann trovò nella tomba a fossa IV tre caldaie in piombo con i bordi rivestiti in bronzo. Le analisi dimostrano che tale

I manufatti di piombo della media e tarda Età del bronzo rinvenuti a Santorino *(in colore)* e a Creta *(in nero)* provengono primariamente dal Laurio. Gli oggetti di Santorino sono del periodo 1550-1500 a.C. Essi comprendono 21 pesi in piombo, chiodi per riparare ceramiche, piombi per reti da pesca, incavi per stipiti di porte da Akrotiri e tre campioni di litargirio *(quadratini pieni)*. Solo un manufatto era in piombo di Sifno. I 22 oggetti rinvenuti a Creta appartengono al periodo 2050-1330 a.C. a Cnosso *(quadratini e rombi vuoti)* e al periodo 1550-1200 a.C. a Kommos *(pallini pieni)*. Solo due manufatti di Cnosso sono di piombo di Sifno. Due sono di metalli di origine sconosciuta, così come uno dei tre pezzi di Kommos, dove erano rari gli oggetti in piombo.

Il piombo e l'argento di cui sono costituiti i manufatti rinvenuti in Attica, nell'Argolide e nella Laconia *(in nero)* e alcuni manufatti egiziani *(in colore)* sono risultati originari del Laurio. I quadratini vuoti illustrano i rapporti fra gli isotopi del piombo in tre caldaie micenee trovate nella tomba a fossa n. IV da Heinrich Schliemann. I pallini vuoti illustrano le proporzioni in quattro pesi in piombo rinvenuti in una tomba a *tholos* a Vapheio. I triangoli vuoti illustrano i rapporti isotopici in tre anelli d'argento e in tre campioni di filo di piombo da Perati; i due triangoli capovolti illustrano le proporzioni in campioni di piombo e di litargirio dell'Elladico medio da Thorikos e i sei rombi indicano i rapporti isotopici in piombo proveniente da Atene. I sei manufatti egiziani *(pallini in colore)* comprendono tre oggetti di piombo della XVIII dinastia, uno d'argento della X dinastia, uno d'argento dell'XI dinastia e uno d'argento della XII o XIII dinastia.

I blocchi di marmo a Demoliaki, nel Laurio, formavano un condotto circolare inclinato, di circa 6 metri, usato per separare il minerale di piombo dalla roccia di scarto più leggera. Le coppelle del condotto trattenevano il minerale pesante quando un flusso d'acqua asportava le scorie.

piombo proviene dal Laurio. Analisi simili di quattro oggetti d'argento trovati nella tomba a fossa IV dimostrano che anche il loro argento proviene dal Laurio.

L'Egitto era famoso nell'antichità per i suoi tesori in oro, ma era del tutto privo di giacimenti locali d'argento. L'argento era di conseguenza molto apprezzato dagli egizi. Una parte dell'oro rinvenuto nelle tombe a fossa di Micene non potrebbe essere provenuto dall'Egitto in cambio di argento del Laurio? Le nostre analisi preliminari di manufatti di piombo e d'argento dell'Egitto dinastico hanno individuato finora tre oggetti di ciascun metallo che presentano la composizione isotopica caratteristica del Laurio. I tre oggetti in piombo sono tutti della XVIII dinastia (1570-1293 a.C.) e furono trovati ad Amarna e ad Abido. Due degli oggetti d'argento sono rispettivamente della X dinastia (2175-2135 a.C.) e dell'XI dinastia (2134-1991 a.C.) e furono rinvenuti a Dendera. Il terzo, trovato ad Abido, risale alla XII o XIII dinastia (1991-1668 a.C.). Benché questi risultati non dimostrino direttamente nulla sull'origine degli oggetti d'oro del tesoro di Micene, essi dimostrano che il piombo e l'argento del Laurio pervennero in Egitto sia prima sia dopo il periodo delle tombe a fossa, così come le nostre analisi dei manufatti delle tombe a fossa dimostrano che i micenei di quel periodo avevano accesso al piombo e all'argento del Laurio.

Prima del nostro lavoro si conosceva assai poco sui luoghi d'origine del piombo e dell'argento usati dai popoli dell'Egeo nell'Età del bronzo. Ora le analisi chimiche, l'analisi isotopica, la datazione col metodo della termoluminescenza e con quello del carbonio 14, l'esplorazione di antiche miniere e la scoperta dei resti dell'antica metallurgia hanno consentito di stabilire un certo numero di fatti. È evidente che le popolazioni insulari del Cicladico primitivo si procuravano questi metalli quasi in parti uguali da Sifno e dal Laurio. In seguito il luogo d'origine dei metalli usati a Tera fu prevalentemente il Laurio, mentre quantità minori continuavano a essere fornite da Sifno. Anche a Creta nell'Età del bronzo recente la fonte principale era il Laurio, anche se una piccola quantità di metallo proveniva da Sifno e da luoghi d'origine non individuati. Nella Grecia continentale gli oggetti finora venuti in luce da Thorikos, Perati, Atene, Vapheio e Micene indicano nel Laurio il luogo d'origine esclusivo, e argento e piombo del Laurio sono stati identificati in Egitto in manufatti della X, XI, XII e XVIII dinastia.

Lo sfruttamento di giacimenti di Sifno nell'Età del bronzo, inferita dalle analisi degli isotopi del piombo contenuti in manufatti, è stato confermato da un'investigazione intensiva delle miniere di Sifno e dalla datazione diretta delle attività minerarie ad Agios Sostis in tale isola. Studi recenti compiuti da Mussche e da Spitaels della miniera n. 3 a Thorikos nel Laurio hanno dimostrato che le attività estrattive vi cominciarono al più tardi nel periodo Antico-elladico III (circa 2700 a.C.). Questa conclusione conferma i dati emersi dalle nostre analisi dei manufatti.

Emerge così un quadro molto semplice: nell'Età del bronzo sia i greci della terraferma sia la popolazione delle Cicladi avevano essenzialmente solo due fonti di piombo e d'argento. Nell'Egeo esistevano molti altri luoghi di produzione di questi due metalli: quali sono dunque le ragioni di questa preferenza accordata al Laurio e a Sifno? È molto probabile che essa si debba al fatto che il metallo ricercato era principalmente l'argento, mentre il piombo era essenzialmente un prodotto secondario. Saggi dei minerali dei siti minerari delle due aree dimostrano che essi contengono il più alto tenore d'argento (circa il 5 per cento in entrambi i casi) fra quelli di tutti i siti minerari dell'Egeo. Se si considera che nei tempi antichi lo sfruttamento delle miniere, la fusione e la coppellazione richiedevano un duro lavoro e grandi quantità di legna, non sorprende che minatori e artigiani preistorici ricercassero i depositi minerali più ricchi.

È stato sostenuto che il piombo e l'argento sarebbero stati introdotti nella civiltà egea dal Mediterraneo occidentale, ma non esiste alcuna prova a sostegno di tale tesi. Molto scarsi sono anche gli indizi a sostegno di una provenienza di argento dall'Asia Minore (finora si sarebbe rinvenuto solo un manufatto nelle Cicladi). Queste conclusioni sono in accordo con l'opinione di Renfrew che nella tarda Età del bronzo il commercio dell'Egeo con paesi stranieri fosse minimo rispetto al flusso interno di merci e materiali.

Il lavoro di Karl M. Petruso, dell'Università di Boston, sui pesi per bilancia che sono stati rinvenuti in grande abbondanza ad Akrotiri e a Zea, e in numero minore in una ventina di siti a Creta, nelle Cicladi e sulla terraferma greca, ha fornito un forte sostegno alla posizione di Renfrew. È stato Petruso a dimostrare che tali pesi, assieme ad alcuni fatti di materiali diversi dal piombo, sono tutti multipli dell'unità-base di 61 grammi; questo sistema di pesi è a quanto pare un sistema minoico sviluppatosi senza alcun apporto dall'esterno dell'Egeo. Un tale sistema avrebbe un senso solo nell'ipotesi che fosse applicato alla valutazione di una qualche merce o gruppo di merci su una base comune in luoghi diversi, con lo stesso sistema.

Il bisogno di un tale sistema comune si comprende a sua volta solo supponendo che esistesse una rete commerciale nell'Egeo e che vi fosse uno scambio pacifico di merci. Recentemente Jack Davis, dell'Università dell'Illinois, ha richiamato l'attenzione su una documentazione archeologica che suggeriva l'esistenza di contatti frequenti, forse regolari, fra Creta, le isole delle Cicladi occidentali (in particolare Santorino, Milo e Zea) e la terraferma (l'Attica e il Peloponneso nord-orientale) a partire almeno dal Minoico medio I sino al Tardo-minoico I A. La documentazione da noi raccolta sull'esistenza di piombo e argento nel Tardo-minoico I, Tardo-elladico I e in tempi posteriori a Creta, Santorino, Zea (dati ottenuti quando abbiamo completato questo articolo) e nei siti della terraferma di Thorikos, Perati, Micene e Vapheio suggerisce che il piombo e l'argento fossero fra le voci principali del commercio lungo questo itinerario occidentale.

GIOCHI MATEMATICI

di *Martin Gardner*

Le ispirate simmetrie geometriche di Scott Kim

Il libro di Scott Kim pubblicato quest'anno da Byte Publications, *Inversions*, è uno dei più deliziosi e sorprendenti che siano mai stati pubblicati. Nel corso degli anni Kim ha sviluppato la magica abilità di rigirare qualsiasi parola, breve frase o lettera in modo tale da far venir fuori una sorta di sorprendente simmetria geometrica. Prendete in considerazione il modo di scrivere il mio nome che vedete sotto, rivoltate il tutto ed, ecco,... rimane esattamente uguale!

Gli esperti in giochi di parole si sono da tempo accorti che si possono formare brevi parole che presentino vari tipi di simmetria geometrica. Sulla Rue Mozart a Parigi una boutique chiamata «New Man» ha una grande insegna su cui è scritto «NeW MaN» con la *e* e la *a* scritte in modo da differire solo per l'orientamento. Il risultato è che l'insegna è simmetrica rispetto al suo rovesciamento. I nomi VISTA (la rivista della United Nations Association), ZOONOOZ (la rivista dello zoo di San Diego) e NISSIN (un'industria giapponese di materiale fotografico) vengono tutti ingegnosamente scritti così da essere simmetrici rispetto al loro rovesciamento.

BOO HOO, DIOXIDE, EXCEEDED, e DICK COHEN DIED 10 DEC 1883 hanno simmetria speculare rispetto a un asse orizzontale. Se li rovesciate e li mettete di fronte a uno specchio, rimangono immutati. Un giorno in un supermercato mia sorella rimase perplessa di fronte al nome scritto su di una scatola di crackers, finché non si accorse che la scatola era stata messa capovolta sullo scaffale. Wallace Lee, un mago della Carolina del Nord, amava divertire gli amici chiedendo loro se avevano mai mangiato degli «ittaybeds», parola che scriveva su un pezzo di carta in questo modo:

ittaybeds

Dopo che tutti avevano risposto di no, aggiungeva: «Per forza, si gustano meglio rovesciati».

Molte parole brevi scritte nei normali caratteri di stampa diventano altre parole se vengono rovesciate. MOM diventa WOW e «do» diventa «op». OSSO rimane uguale. Ecco un modo divertente per scrivere «minimum» in modo che resti uguale se viene ruotato di 180 gradi:

È stato Kim a portare questa curiosa arte della calligrafia simmetrica a livelli mai raggiunti prima. Kim, distorcendo ingegnosamente le lettere (ma neppure tanto da rendere irriconoscibili una parola o una frase), ha prodotto delle configurazioni semplicemente fantastiche. Il suo libro è una raccolta di tali meraviglie, inframmezzate da osservazioni provocatorie sulla natura della simmetria, i suoi aspetti filosofici, il suo inserimento nell'arte, nella musica e nei giochi di parole.

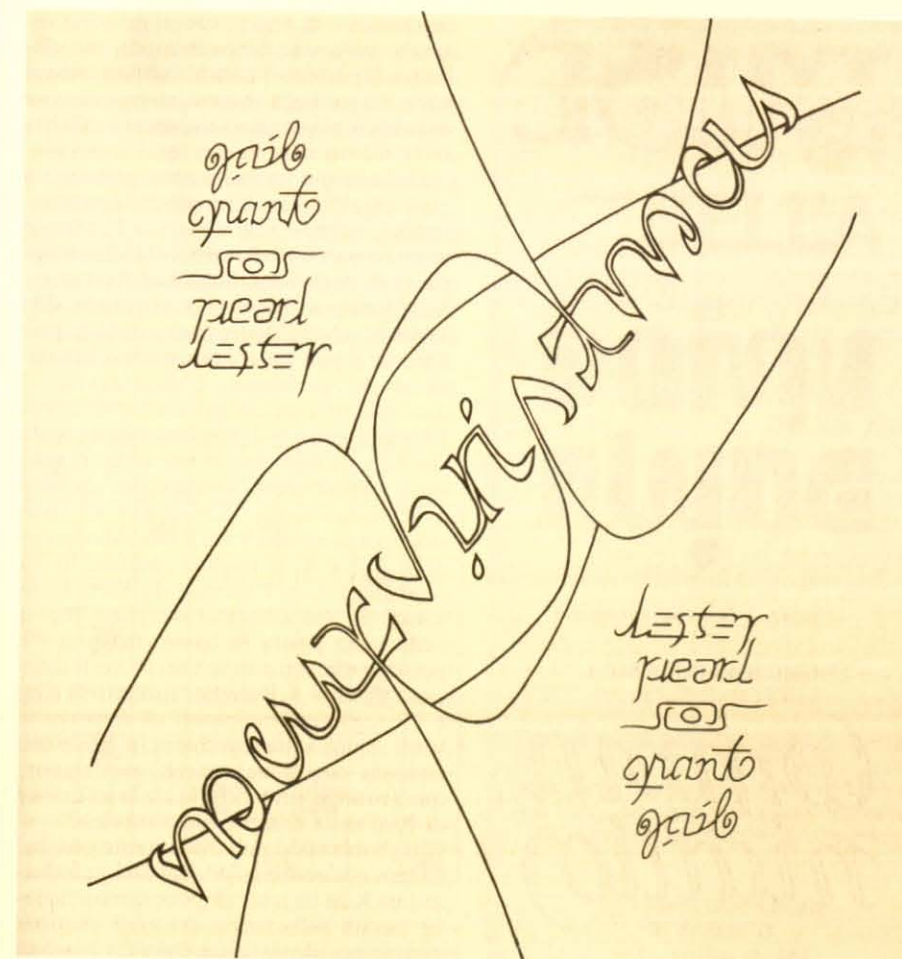
Kim non è uno sconosciuto per questa rubrica: è un giovane di origini corcane, nato negli Stati Uniti, che si sta specializzando in informatica alla Standford University. Era ancora un ragazzo quando ha incominciato a creare giochi matematici molto originali, alcuni dei quali sono comparsi in queste pagine. Tra questi: «viaggio del re smarrito» (agosto 1977), il problema della disposizione dei cavalli (negli scacchi) sui vertici di un ipercubo (giugno 1978), la sua soluzione dell'«inscatolamento della scatola» (giugno 1979) e la sua mappa splendidamente simmetrica (giugno 1980). Oltre ad avere una notevole capacità di pensiero geometrico (non solo in due o tre dimensioni, ma anche nello spazio a quattro e in spazi superiori), Kim è un pianista di musica classica che per anni è stato indeciso se proseguire nello studio della matematica o in quello della musica. Al momento si interessa di come servirsi del calcolatore per progettare caratteri di stampa, campo questo in cui fu pioniere il suo amico e mentore Donald E. Knuth.

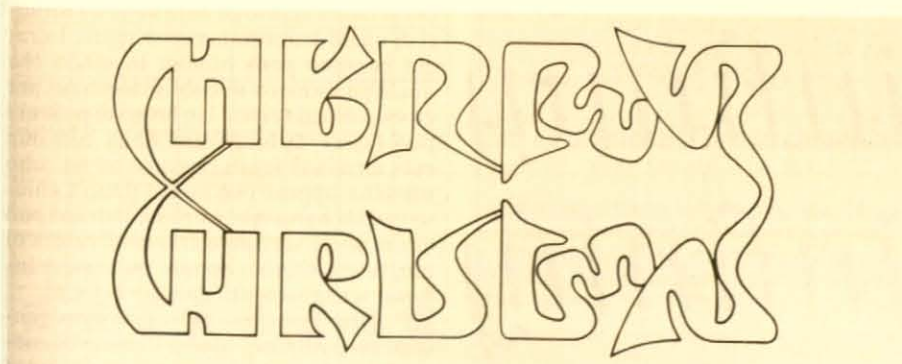
Per diversi anni il talento di Kim per scrivere le parole in modo da dar loro inattese simmetrie venne sfruttato solo per divertire gli amici e per elaborare cartoline natalizie per la famiglia. Quando incontrava uno sconosciuto o una sconosciuta ad una festa, imparava il suo nome, quindi spariva per un attimo e ritornava con il nome graziosamente disegnato in modo che si potesse leggere ugualmente anche rovesciato. Nella figura della pagina a fronte compare il suo biglietto di Natale del 1977 con simmetria di rovesciamento. (Lester e Pearl sono suo padre e sua madre; Grant e Gail suo fratello e sua sorella.) L'anno successivo trovò il modo per rendere «Merry Christmas 1978» specularmente simmetrico con asse di simmetria orizzontale, mentre nel 1979 lo rese con asse di simmetria verticale. Per l'anniversario di matrimonio dei suoi genitori, Kim ideò una torta con una glassa di cioccolata e vaniglia che formava la combinazione che si vede nella pagina a fianco. («Lester» è in nero, «Pearl» è rovesciato, in bianco.) Questa è la tecnica di Kim «figura e sfondo». Ne troverete un altro esempio in *Gödel, Escher, Bach: An Eternal Golden Braid*, il libro vincitore del premio Pulitzer scritto da quel buon amico di Kim che è Douglas R. Hofstadter, con il quale mi alterno quest'anno su queste colonne. A proposito di Kurt Gödel, J. S. Bach e M. C. Escher, in un'altra figura della pagina a fronte si può vedere come Kim abbia dato a ciascun nome un'incantevole simmetria speculare. In una figura a pagina 118 Kim ha scritto l'intero alfabeto in modo tale che la configurazione complessiva ha una simmetria bilaterale.

La calligrafia magica di Kim fu notata da Scot Morris, uno dei redattori di «Omni», che dedicò una pagina della sua popolare rubrica di giochi al lavoro di Kim nel numero di settembre 1979 di «Omni» e bandì un concorso tra i lettori per la ricerca di configurazioni analoghe. Kim venne assunto per giudicare le migliaia di proposte arrivate. Troverete quelle vincenti sul numero di «Omni» dell'aprile 1980 e le migliori classificate nella rubrica di Morris di maggio e del novembre dello stesso anno.

Tutte le configurazioni del libro di Kim sono fatte da lui; eccovene alcune altre a pagina 118 per darvi un'idea della sorprendente varietà di trucchi visivi che Kim estrae dal suo cappello.

Passiamo ora a due insoliti problemi matematici proposti da Kim, entrambi i quali sono tutt'ora solo parzialmente risolti. Nel 1975, quando era alla scuola superiore, Kim escogitò la seguente generalizzazione del vecchio problema di come disporre sulla scacchiera otto regine in modo che nessuna sia sotto attacco. Chiediamoci, si disse Kim, qual è il nume-

Il nome dell'articolista si legge anche capovolto

La cartolina augurale di Kim si legge ugualmente capovolta

«Merry Christmas» è specularmente simmetrico con asse di simmetria verticale

«Merry Christmas» è specularmente simmetrico con asse di simmetria orizzontale

«Lester» e «Pearl» fanno da figura e da sfondo

Gödel, Escher e Bach in simmetria speculare

L'alfabeto specularmente simmetrico

«True» è inserito in «False»

«Upside» è lo stesso rovesciato

Una «comunication» divisa può essere invertita

«Man» e «Woman» invertibili

«Infinity» va all'infinito

ro massimo di regine che si possono disporre sulla scacchiera in modo che ciascuna regina ne attacchi esattamente n altre. Come negli scacchi, stabiliamo che nessuna regina ne possa attaccare un'altra se ce n'è un'altra in mezzo.

Quando n è uguale a zero, abbiamo il problema classico. Kim riuscì a dimostrare che, quando n è uguale a 1, 10 è il numero massimo di regine. (La dimostrazione si trova nel «Journal of Recreational Mathematics» vol. 3, n. 1 pag. 61; 1980-81.) Nella figura in alto della pagina a fronte si può vedere una piacevole soluzione del problema. L'illustrazione al centro mostra una soluzione massimale di 14 regine con $n = 2$, combinazione questa che Kim ha descritto in una lettera come «così orribilmente asimmetrica da non aver nessun diritto di esistere». Ci sono solo ipotesi su quale sia il massimo quando n è 3 o 4. Il miglior risultato di 16 regine per $n = 3$ trovato da Kim, ha la soluzione ridicolmente semplice che si vede nella figura in basso, ma non c'è nessuna dimostrazione che 16 sia il massimo. Per $n = 4$, il miglior risultato di Kim è 20 regine. Siete capaci di disporre le venti regine sulla scacchiera in modo che ciascuna regina ne attacchi esattamente quattro altre, prima che io dia la soluzione di Kim nella mia prossima rubrica?

Il problema si può ovviamente generalizzare a scacchiere di qualsiasi grandezza, ma Kim ha una semplice dimostrazione basata sulla teoria dei grafi che, per quanto grande sia la scacchiera, n non può avere un valore superiore a 4. Per $n = 1$ Kim ha dimostrato che il massimo numero di regine non può superare il massimo intero minore o uguale a $4k/3$, dove k è il numero dei quadrati lungo il lato di una scacchiera. Per $n = 2$, Kim ha una dimostrazione più difficile del fatto che il numero massimo di regine non può superare $2k$-2 e che questo massimo si può ottenere su tutte le scacchiere di ordine pari.

Il problema di Kim riguardante il drago di policubi non era stato ancora reso noto ed io e lui saremo ben lieti se qualche lettore ci darà qualche chiarimento su di esso. Per prima cosa dobbiamo definire un drago come una catena semplicemente connessa di cubi unitari identici congiunti per le facce in modo che ciascun cubo (tranne il cubo alla fine della catena) sia attaccato ad esattamente due altri cubi. Il drago può piegarsi in qualsiasi direzione possibile purché nessun cubo interno tocchi la faccia di nessun altro cubo all'infuori di quelli immediatamente vicini. I draghi possono però ruotare in modo che qualsiasi numero di cubi si tocchino per gli spigoli o ai vertici. Un drago di policubi può essere di lunghezza finita con due cubi terminali legati ciascuno ad un cubo soltanto, oppure può essere finito e chiuso così da non avere estremi. Il drago può anche avere un solo estremo ed essere di lunghezza infinita, oppure può essere infinito senza nessuno dei due estremi.

Poniamo ora una domanda apparentemente semplice: qual è il minor numero di draghi necessario per riempire tutto lo spazio? Possiamo porre la questione in un altro modo. Immaginiamo che lo spazio sia completamente riempito da un numero infinito di cubi unità. Qual è il più piccolo numero di draghi in cui si può dividere tagliando lungo i piani che definiscono i cubi?

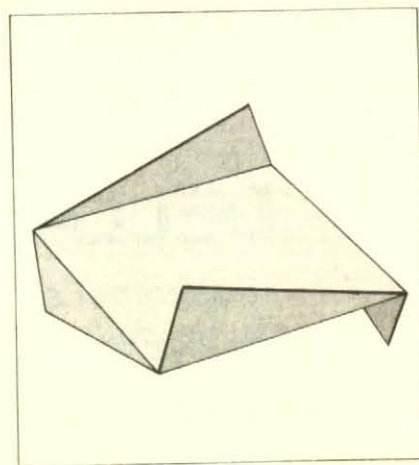
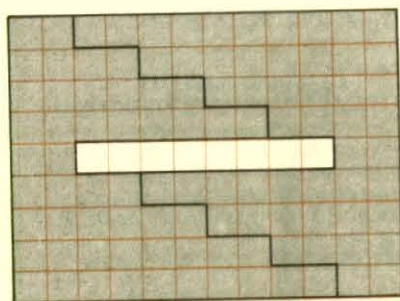
Se consideriamo il corrispondente problema bidimensionale (draghi composti da quadrati unità), è facile rendersi conto che la risposta è due. Basta semplicemente intrecciare due spirali di draghi piani infiniti con un estremo, uno grigio e uno bianco, come si può vedere a sinistra in alto nella pagina seguente.

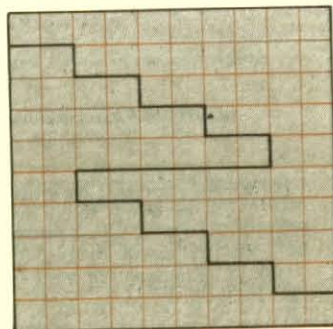
n = 1

n = 2

n = 3

Il problema delle regine di Kim

Tassellatura del piano con due draghi piani

Foglio con simmetria speculare di rotazione

La coperta patchwork rovinata viene ricucita così da formare una coperta quadrata senza buchi

Come si può suddividere un quadrato in nove triangoli acuti

Riempire lo spazio tridimensionale con draghi di policubi non è così semplice. Kim ha trovato un modo per piegare quattro draghi di lunghezza infinita ad un estremo (conviene immaginarli di diverso colore) in una struttura di forme elicoidali intrecciate che riempie tutto lo spazio. Il sistema è troppo com
venir spiegato in uno spazio limitato; vi basti la mia parola che è possibile.

Si può fare con tre draghi? Non solo questa domanda non ha risposta, ma lo stesso Kim non è riuscito a dimostrare che non si può fare con due draghi! Ha scritto in una lettera: «Una soluzione con due soli draghi rappresenterebbe una sorta di simbolo yin-yang tridimensionale infinito: lo spazio negativo lasciato da un drago sarebbe l'altro drago. È la bellezza di questo intreccio e la possibilità di costruire un modello grande abbastanza da muoversi che mi fa continuare a cercare una soluzione».

Il problema si può ovviamente generalizzare a draghi fatti di cubi unità di un qualsiasi numero di dimensioni. Kim ha ipotizzato che in uno spazio di n dimensioni il più piccolo numero di draghi che lo riempia completamente è $2(n-1)$, ma la cosa è molto incerta.

Alcuni anni fa ebbi il piacere di spiegare il problema del drago di policubi al matematico di Cambridge John Horton Conway. Quando conclusi dicendo che Kim non aveva ancora dimostrato che due draghi non possono tassellare lo spazio tridimensionale, Conway disse immediatamente: «Ma è ovvio che...». Si bloccò a metà frase, fissò per un minuto lo spazio e poi esclamò: «*Non* è ovvio!

Non ho idea di quello che sia passato nella mente di Conway in quel momento. Posso solo dire che se l'impossibilità di riempire lo spazio tridimensionale con due draghi non è sembrata ovvia a Conway o a Kim, probabilmente non risulterà ovvia a nessun altro.

Ecco le soluzioni ai brevi problemi di giugno:

1. Si considerino i due punti più esterni A e K. Un qualsiasi punto L sulla linea tra A e K (compresi) avrà la stessa somma di distanze da A e K. Consideriamo ora B e J, la prima coppia di punti che si incontra procedendo verso l'interno. Come prima, per minimizzare la somma delle distanze di L da B e J, L deve essere tra B e J. Dato che L è anche tra A e K, la sua posizione renderà minima la somma delle distanze da A, K, B e J.

Continuate in questo modo prendendo le coppie di punti che si incontrano verso l'interno. L'ultima coppia di punti è E e G. Tra di loro c'è il singolo punto F. Qualsiasi punto tra E e G minimizzerà le distanze rispetto a tutti i punti tranne F. Ovviamente, se volete rendere minima anche la distanza da F, bisogna che il punto sia esattamente in F. In breve Lavinia dovrebbe spostarsi nello stesso palazzo dove vive il suo amico Frank.

Generalizzando, per un qualsiasi numero pari di punti su una linea retta, qualsiasi punto tra i due punti intermedi avrà la somma minima di distanze da tutti i punti. Per un qualsiasi numero dispari di punti, la posizione centrale è la posizione desiderata. Il problema è comparso su *No Calculus, please* di J. H. Butchard e Leo Moser («Scripta Mathematica» vol. 18, nn. 3-4, pagg. 221-236; settembre-dicembre, 1952).

2. Nella figura in alto a destra della pagina a fronte si può vedere come piegare un quadrato di carta in una forma che non ha piano di simmetria, ma che può tuttavia essere
magine speculare. Si dice che la figura ha una simmetria speculare di rotazione, un tipo di simmetria di grande importanza in cristallografia. Ho tratto questo caso dalla pag. 42 di *Simmetry in Science and Art* di A. V. Shubnikov e V. A. Koptsik (Plenum Press,

3. L'illustrazione al centro della pagina a fronte mostra come si possa tagliare la coperta patchwork rovinata in due parti da cucire insieme per formare una coperta quadrata senza buchi. Si tratta del problema n. 215 di *Puzzles and Curious Problems* di Henry Ernest Dudeney (Thomas Nelson and Sons, Ltd., 1931).

4. L'illustrazione in basso nella pagina a fronte mostra come tagliare un quadrato in nove triangoli acuti. Se si considera la triangolazione nel suo significato topologico, non si ammette cioè che un vertice si trovi su un lato di un triangolo, allora non esiste una soluzione con nove triangoli benché ce ne sia una per otto triangoli, per 10 e per tutti i numeri superiori a 10. Questo curioso risultato è stato dimostrato in un articolo non pubblicato da Charles Cassidy e Graham Lord della Laval University del Quebec.

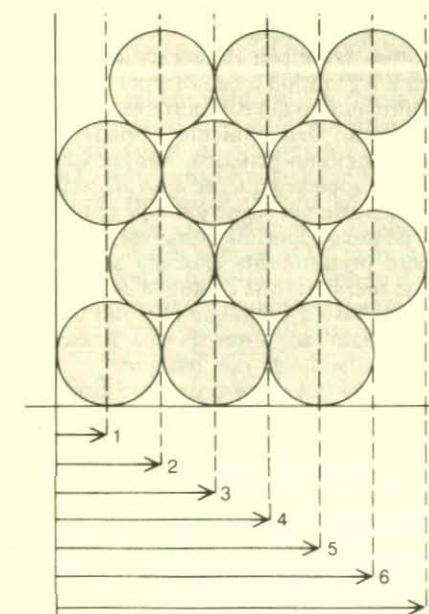
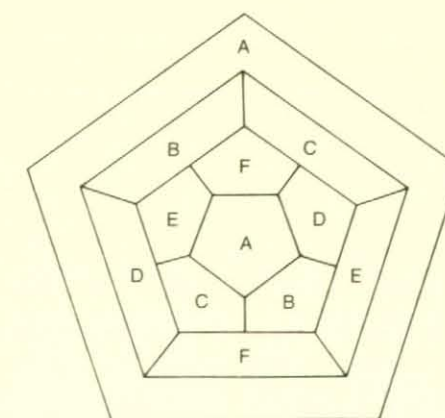
Nella figura in alto vengono presentati quattro modi in cui tagliare un triangolo equilatero in cinque triangoli isosceli. Nella prima combinazione non ci sono triangoli equilateri, nella seconda e nella terza ce n'è uno e nella quarta due. Le quattro combinazioni, ideate da Robert S. Johnson, compaiono in «Crux Mathematicorum» (vol. 4, n. 2, pag. 53; febbraio 1978).

5. Le distanze intere su una retta si possono misurare in centimetri con una moneta giapponese da uno yen con raggio di un centimetro, nel modo che si vede nella figura in basso a sinistra.

6. La mappa della figura in basso a destra trovata da Lloyd Shapley, matematico della Rand Corporation, dimostra che, quando nel gioco della colorazione delle mappe di Steven J. Brams si usano cinque colori, c'è una mappa sulla quale il massimizzatore può sempre vincere.

La mappa è una proiezione dello scheletro di un dodecaedro, con una regione esterna *(A)* che rappresenta la faccia posteriore del solido. La strategia del massimizzatore è quella di giocare sempre sulla faccia del dodecaedro opposta a quella su cui ha fatto l'ultima giocata l'avversario, servendosi del suo stesso colore. (Nella figura le regioni che rappresentano coppie di facce opposte sono contrassegnate con la stessa lettera). Come potete facilmente vedere, questa strategia impedisce che si possano usare ancora colori successivi e fa sì che la partita arrivi al punto in cui il minimizzatore deve servirsi di un sesto colore per giocare.

7. La strategia di John Horton Conway per il suo gioco del whim è la seguente. Muovete come se ci fosse un altro mucchio di gettoni contenente un solo gettone, se c'è un mucchio di quattro o più gettoni, come se ci fosse un altro mucchio di due gettoni, se non c'è nessun mucchio con quattro o più gettoni. Finché qualcuno non fa la mossa del whim, il mucchio invisibile rimane, mentre scompare quando qualcuno compie la mossa del whim. Una posizione è vincente nel whim per il giocatore che vincerebbe nel gioco nim se ci fossero effettivamente i mucchi del whim. La strategia vincente, quindi, consiste semplicemente nell'immaginare che ci sia il whim finché un gioca
lo elimina e nel giocare la strategia del nim normale. Se muovendo si passa da una posizione in cui almeno un mucchio ha quattro o più gettoni ad una in cui non ce n'è nemmeno uno, la pila del whim guadagna il suo secondo gettone dopo la mossa, non prima.

Quattro modi per suddividere un triangolo equilatero in cinque triangoli isosceli

Monete da uno yen per misurare distanze intere

Dimostrazione di come si può vincere un gioco di colorazione delle mappe